2019

Automatic ^{13}C Chemical Shift Reference Correction of Protein NMR Spectral Data Using Data Mining and Bayesian Statistical Modeling

Xi Chen
University of kencutky, billchenxi@gmail.com
Author ORCID Identifier:
https://orcid.org/0000-0001-7094-6748
Digital Object Identifier: https://doi.org/10.13023/etd.2019.057

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Xi Chen, Student

Dr. Hunter Moseley, Major Professor

Dr. Trevor Creamer, Director of Graduate Studies

Automatic ^{13}C Chemical Shift Reference Correction of Protein NMR Spectral Data Using Data Mining and Bayesian Statistical Modeling

---

DISSERTATION

A dissertation submitted in partial fulfillment of the

requirements for the degree of Doctor of Philosophy in the

College of Medicine

at the University of Kentucky

By

Xi Chen

Lexington, Kentucky

Director: Dr. Hunter Moseley

Professor of Molecular and Cellular Biochemistry

Lexington, Kentucky

2019

https://orcid.org/0000-0001-7094-6748

ABSTRACT OF DISSERTATION

Automatic ^{13}C Chemical Shift Reference Correction of Protein NMR Spectral Data Using Data Mining and Bayesian Statistical Modeling

Nuclear magnetic resonance (NMR) is a highly versatile analytical technique for studying molecular configuration, conformation, and dynamics, especially of biomacromolecules such as proteins. However, due to the intrinsic properties of NMR experiments, results from the NMR instruments require a refencing step before the down-the-line analysis. Poor chemical shift referencing, especially for ^{13}C in protein Nuclear Magnetic Resonance (NMR) experiments, fundamentally limits and even prevents effective study of biomacromolecules via NMR. There is no available method that can rereference carbon chemical shifts from protein NMR without secondary experimental information such as structure or resonance assignment.

To solve this problem, we constructed a Bayesian probabilistic framework that circumvents the limitations of previous reference correction methods that required protein resonance assignment and/or three-dimensional protein structure. Our algorithm named Bayesian Model Optimized Reference Correction (BaMORC) can detect and correct ^{13}C chemical shift referencing errors before the protein resonance assignment step of analysis and without a three-dimensional structure. By combining the BaMORC methodology with a new intra-peaklist grouping algorithm, we created a combined method called Unassigned BaMORC that utilizes only unassigned experimental peak lists and the amino acid sequence.

Unassigned BaMORC kept all experimental three-dimensional HN(CO)CACB-type peak lists tested within ± 0.4 ppm of the correct ^{13}C reference value. On a much larger unassigned chemical shift test set, the base method kept ^{13}C chemical shift referencing errors to within ± 0.45 ppm at a 90% confidence interval. With chemical shift assignments, Assigned BaMORC can detect and correct ^{13}C chemical shift referencing errors to within ± 0.22 at a 90% confidence interval. Therefore, Unassigned BaMORC can correct ^{13}C chemical shift referencing errors when it will have the most impact, right before protein resonance assignment and other downstream analyses are started. After assignment, chemical shift reference correction can be further refined with Assigned BaMORC.

To further support a broader usage of these new methods, we also created a software package with web-based interface for the NMR community. This software will allow non-NMR experts to detect and correct ^{13}C referencing errors at critical early data analysis steps, lowering the bar of NMR expertise required for effective protein NMR analysis.

KEYWORDS: NMR, referencing correction, statistical model, protein.

Xi Chen

(Name of Student)

03/12/2019

Date

Automatic ^{13}C Chemical Shift Reference Correction of Protein NMR Spectral Data Using Data Mining and Bayesian Statistical Modeling

By

Xi Chen

Hunter Moseley

Director of Dissertation

Trevor Creamer

Director of Graduate Studies

03/12/2019

Date

DEDICATION

To my mom, thank you for teaching me to have hope and be brave.

To all my fans. Thanks for being there for me! I love y'all!!

To Hunter Moseley. Thanks for the opportunity.

ACKNOWLEDGMENTS

The following dissertation, while an individual work, benefited from the insights and direction of several people. First, my Dissertation Chair, Dr. Hunter Moseley, without his support, all of this will not be possible. In addition, many professors of mine provided timely and instructive comments and evaluation for dissertation process, allowing me to complete this project on schedule. Next, I wish to thank the complete Dissertation Committee, and outside reader, respectively: Dr. Peter Spielmann, Dr. Konstantin Korotkov, Dr. Arny Stromberg, and Dr. Jerzy Jaromczyk. Each individual provided insights that guided and challenged my thinking, substantially improving the finished product.

In addition to the technical and instrumental assistance above, I received equally important assistance from family and friends. Andrey Smelter provided on-going support throughout the dissertation process, as well as technical assistance critical for completing the project in a timely manner. Finally, I wish to thank everyone (I couldn't list all of you otherwise these acknowledgements will be too long) who have helped me and supported me through the journey of the grad school. From the bottom of my heart, thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 Protein NMR reference correction

Since its discovery in the work of Rabi[1], Purcell[2] and Bloch[3], Nuclear magnetic resonance (NMR) has developed into a highly versatile and widely used analytical technique for the study of molecular configuration, conformation, and dynamics, especially of biomacromolecules such as proteins[4-12]. The typical NMR experiment is commonly divided into four important stages: (1) sample preparation, (2) spectroscopy, (3) raw data processing, and (4) analysis. Each stage of the experiment contributes to the success of any NMR experiment, and if any step is ignored or improperly implemented, the whole experiment can be doomed.

After nearly eight decades of evolving, almost every aspect of the NMR experiment has been drastically improved. Sample preparation in solid-state allows the advantage of NMR on large biomolecular assemblies such as an intact virus. In the 1960s an NMR spectrometer of 60 MHz (1.4T) was considered state of the art[13], while the emerging of hybrid magnet that allow NMR experiment operating at 1500 MHz (35.2T) in the solid state could be the new high-field standard soon[14]. Down-the-line analysis tools for NMR research also caught up. From the raw data obtain from the spectrometer through a variety of mathematical operations, e.g. digital filtering of solvent, apodization etc. prior to Fourier transformation, to the subsequent analysis of the processed data, e.g. resonance assignment and of the extraction of constraints for generating of atomic models, all have gone through certain improvements and innovations. However, one crucial step in data processing, spectral referencing, which is performed after collection of the raw data in vast

majority of NMR experiments, and before the analysis such as assignment, hasn't changed much.

Several factors are fundamental to the utilization of NMR spectral data: resonance sensitivity, spectral precision, and spectral accuracy[15,16]. While various improvements in sample preparation[17,18], instrumentation[19-22], and pulse sequences[23,24] have greatly improved resonance sensitivity and spectral precision, spectral accuracy still depends on the same basic procedure: referencing chemical shifts to a designated chemical standard.

Additionally, variance in chemical shifts can be caused by a variety of experimental factors, including pH, temperature, presence of salts, and use of organic solvent mixtures. These factors along with simple human error can lead to inaccurate referencing[25,26]. In protein NMR analyses, 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) is the recommended internal standard for chemical shift referencing[27,28]among two other commonly used options, trimethylsilyl propanoic acid (TSP) and 4,4-dimethyl-4-silapentane-1-ammonium (DSA).

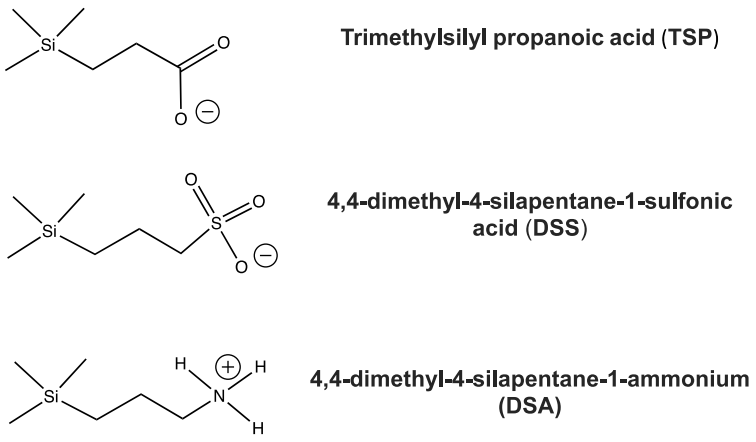

Figure 1.1 Recommended internal references.

However, DSS has a negative charge within NMR-relevant pH ranges and can interact with positively charged residues of a protein of interest, altering its reference chemical shift value [25.] Additionally, temperature affects the reference chemical shift of

DSS, requiring a temperature correction step in DSS-based referencing. The general procedures is as follows: before the NMR experiment, the sample of interest will be carefully doped with a small amount (~50 μM) of an internal reference, typically DSS[27]. Lack of experience with chemical shift referencing and the factors that can affect referencing is a major contributor to chemical shift referencing inaccuracy. All downstream analyses and interpretations are affected by these inaccuracies in chemical shifts, including the assignment of resonances in biomacromolecules such as proteins. Moreover, these inaccuracies can outright prevent data analysis, especially with semiautomated data analysis tools, or propagate through data analysis, snowballing into interpretive errors with respect to structure and dynamics. Since the structural and dynamic information contained in the chemical shift is subtle, even small chemical shifts errors due to inaccurate referencing may provide a distorted representation of the protein, especially when chemical shifts are directly used in structure determination[18,20-22].

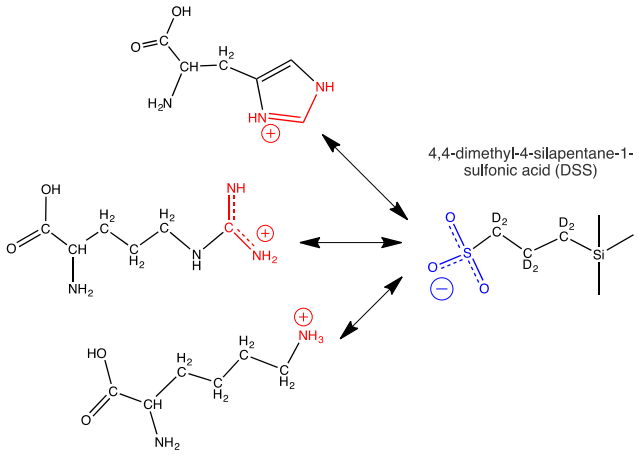

Figure 1.2 Interaction between protein and DSS. Negative charges of the amino acid residue, this could lead to shift where the DSS peak location, and lead to reference errors.

1.2 Motivation

To address these issues in protein NMR, we have developed a new methodology referred to as Bayesian Model Optimized Reference Correction (BaMORC), which detects

and corrects ^{13}C chemical shift referencing errors using sets of C_α and C_β chemical shift pairs. BaMORC minimizes the difference between the known amino acid frequencies based on the protein sequence and the frequencies predicted using a set of bivariate statistical models that are amino acid and secondary structure specific [27] and are based on C_α and C_β chemical shift statistics. The minimization comes from the adjustment of the ^{13}C chemical shift referencing. The statistical models integrate prior amino acid and chemical shift propensity information along with amino acid and secondary structure probabilities calculated using a chi-squared statistic based on C_α and C_β chemical shifts and refined chemical shift statistics derived from the RefDB. The refined expected values, variances, and covariances for C_α and C_β chemical shifts are derived from 1557 RefDB assigned chemical shift entries that were selected and filtered using a variety of statistically driven data mining methods. Since RefDB only includes datasets from proteins with well-defined structure, BaMORC is likewise tuned to work with chemical shift datasets from proteins with well-defined structure. We integrated BaMORC with a new intra-peak list grouping algorithm[28] developed in our laboratory to create a combined method, which we refer to as Unassigned BaMORC , that can correct ^{13}C chemical shift referencing using an unassigned three-dimensional HN(CO)CACB-type peak list[29]. Thus, ^{13}C chemical shift referencing can be automatically analyzed and corrected before downstream analyses, including protein resonance assignment. Unassigned BaMORC generates a correction value, a file of re-referenced chemical shifts and a residual plot, which shows the optimization of the predicted amino acid frequencies and the point at which the best reference correction value occurs in the optimization. Furthermore, we have implemented

an Assigned BaMORC method that can utilize assigned chemical shifts to improve reference correction after resonance assignment.

1.3 Dissertation outline

Chapter 2 reviews the important background with respect to the biological application of protein NMR. It further explains the importance of referencing and the problem related to referencing that NMR community is currently facing. At the end of chapter 2, we provide a general, high-level description of the solutions to the referencing problem implemented in our Bayesian Model Optimized Reference Correction (BaMORC) method.

Chapter 3 provides the general design principles and fundamental statistical background of the methodology. It explains the statistics inference from data-driven approach and the "nuts and bolts" of the BaMORC statistical model.

Chapter 4 describes the project design overview, from the data collection, data cleaning, and core algorithmic components including optimization.

Chapter 5 describe the heart of the dissertation—BaMORC, a tool for protein NMR reference correction. This chapter provides details on the integration of the algorithmic pieces described in Chapter 4 into the BaMORC algorithm, as well as its development, performance, and limitations.

Chapter 6 focuses on the BaMORC package, which represents the practical implementation of Chapter 5. Besides introducing the functionality of the BaMORC package, which including the command-line interface, I further documented how to setup the program running environment and installation of the package, in hope to provide a one-stop shop for readers who are interested in using the package.

To further extrapolate the usage of the BaMORC and to offer a user-friendly access to a broader audience who are not familiar with NMR technology but still want to analysis chemical shift data, Chapter 7 describes the BaMORC web application.

Chapter 8 summarizes this dissertation and includes a discussion of the context of the research as well as future directions of the project. Beyond this dissertation, several improvements and features could be included in the BaMORC analysis project to broaden its application.

CHAPTER 2. PROTEIN NMR HISTORY AND ITS BACKGROUND

2.1 NMR history

The fundamental purpose of an NMR spectrometer is to measure the frequency of the resonance for particular nuclei. The basic nuclear magnetic resonance relationship was established as the Larmor equation (Equation 1.1) during the discovery of this physical phenomenon[30]. Equation $\omega = \gamma B$ suggested the resonance frequency of a nucleus (ω) equals to the magnetogyric ration (γ, specific to nucleus type) times the external magnetic field (B).

At the beginning, it was thought the frequency of a nucleus depended entirely on the strength of the magnetic field that was present. Only after development in the stability and homogeneity of the magnetic field, and after three separate resonances of hydrogen atoms in ethanol were observed, this phenomenon eventually became known as "chemical shift" as the frequency of a resonating nucleus is largely dependent on the local chemical environment surrounding the nucleus[31]. Later as the resolution improved, separated peaks of same resonance can be observed as single peak lines, and this improvement contributed the discovery of the concept of indirect spin-spin coupling[32].

Around the late 1950s and early 1960s, as the strength of the magnetic field increased to over 100 MHz and the emergence of instruments that allow a constant relationship between the field and radiofrequency (RF), the NMR spectrum scan collection time dramatic decreased to a constant. With the advent of double resonance, carbon spectroscopy eventually overcame its limitation of the low ^{13}C natural abundance[33], since applying two RF fields simultaneously to a sample allows the measurement of one spin system while the other is perturbed. Experiments with spin decoupling methods and the

ability to detect the nuclear Overhauser effect were then introduced, providing NMR features sensitive to molecular conformation[34].

Ernst and Anderson showed in 1966 in their work that a Fourier Induction Decay (FID) following a short RF pulse was enough to produce a spectrum from a range of frequencies, and Fourier transform NMR (FT-NMR) was developed with the aid of a computational interface directly to the spectrometer[35]. These innovations revolutionized NMR spectroscopy through a large decrease in collection times, thus improving spectrometer sensitivity, which is the main disadvantage of the NMR as compared with other spectroscopy techniques.

Over the recent 40+ years, other improvements such as pulse sequences further advanced NMR spectroscopy. Right now, NMR has become one of the most versatile analytical tools for detecting and characterizing molecular phenomena across many fields of research. For example, Magnetic Resonance Imaging (MRI), an application of NMR with extremally wide bore magnets, allows studying sample in vivo (even human and large animal) via the measurement of frequencies across spatial gradient[36].

Today, NMR as a standard and must-have research instrument allows studies ranging from small (metabolite) to large (protein) biomolecules in a variety of experimental environments (solution, solid or complex mixtures).

2.2 Protein NMR

Assignment of resonances in NMR spectra of a given protein is the first step in any NMR study that is interested in protein structure, structural interactions, and dynamics. After the introduction of the correlated spectroscopy (COSY) and nuclear Overhauser effect spectroscopy (NOESY) around the 1980s, NMR spectral resolution dramatically

improved[37,38]. These two-dimensional (2D) NMR techniques allowed the development of systematic methods of assignment that relied only on protein sequence information, i.e. sequential assignment methodology[39,40]. With the development of three-dimensional (3D) ^{15}N-edited NMR method in late 1980s, the size of protein can be studied using sequential assignment method increased and resolution further improved[41,42]. In the 1990s, through-bond scalar coupling assignment for doubly ^{13}C and 15N isotopic labeled proteins facilitates the study of even larger biomacromoleculer systems[43].

Combining this triple resonance method with deuteration and transverse relaxation optimized spectroscopy (TROSY), the mass limit was later pushed beyond 40 kD[44], and a complete assignment for ^{1}H, ^{13}C and ^{15}N can be obtained from double labelling and triple resonance assignment NMR experiment. Recent progress in magic-angle spinning (MAS) solid-state (SS) NMR techniques has enable the studies of large, unoriented membrane proteins with a huge jump on the weight size limitation of 114 kD[45]. Also, Lewis Kay has demonstrated in his papers that solution NMR has the capability to analyze macromolecular complexes over 500 kD[46] through the use of selectively deuterated sample production methods[47].

While structure determination is often the focus of protein NMR research, there are many different types of structural and interaction information provided from a wide variety of NMR experiments. Of particular interest to this research, information about protein secondary structure is provided by ^{1}H, ^{13}C and ^{15}N chemical shifts and backbone φ and ψ torsion angles[47-50]. This secondary structure information can be used for NMR referencing correction[51], which will be introduced in the next section.

2.3 Protein NMR referencing

A core type of information provided by almost all NMR experiments are chemical shifts. Chemical shift is a principle NMR feature used for resonance assignment and analysis of NMR data. The origin of the chemical shift is the characteristic variation of resonant frequency, or Larmor frequency (υ), between each type of nucleus gyromagnetic ratio (γ) due to the difference of a given external magnetic field. And the Larmor frequency υ is calculated as $\upsilon = -\gamma B$. As illustrated in section 1.1, many factors could shackle those four main stages of an NMR experiment, leading to inaccurate or imprecise chemical shifts values. One fundamental and inevitable factor that contributes to this issue is inaccurate referencing due to human error[25,26].

The reason why chemical shifts require a reference is that the chemical shift, a resonant frequency of a nucleus in a magnetic field, is a relative measurement instead of an absolute measurement. This value is calculated from reference frequency $\delta = \frac{\nu_{sample} - \nu_{ref}}{\nu_{ref}}$, where ν_{sample} and ν_{ref} are the absolute resonance frequency of the sample and of a standard reference respectively. Therefore, inaccurate referencing will contribute to mis-assignment and could eventually lead to unrealistic interpretations of NMR data, including protein structural errors. With a small deviation of ^{13}C chemical shift measurements on the order of 0.3 ppm (0.05 ppm for ^{1}H and 0.5 ppm for ^{13}N) can lead to mis-identification of secondary structure[52-57].

Three major referencing methods are used to reference chemical shifts in a protein NMR experiment: 1) internal referencing, 2) external (substitution) referencing, and 3) statistical modeling approach. The most common referencing method, internal referencing, is performed through the addition of an internal standard directly (internally) into the

sample under study[58,59], but this approach will contaminate the sample and may affect the chemical shifts of interest. External (substitution) referencing involves separating the sample and reference with a glass wall, either in same or different tubes without sample contamination; however, any magnetic susceptibility differences introduced by the glass or physical separation need to be corrected theoretically[58]. A statistical modeling approach uses statistics extracted from NMR database to estimate the reference value based on a statistical model; however, the performance of this approach is heavily dependent on the model's representative performance[51,60,61].

Solution NMR protein chemical shifts are normally referenced to an internal standard that is soluble in the NMR sample. Commonly used internal standards for protein NMR experiments include Trimethylsilyl propanoic acid (TSP), 4,4-dimethyl-4silapentane-1-sulfonic acid (DSS) and 4,4-dimethyl-4-silapentane-1-ammonium (DSA) as shown in Figure 1.2. Among these three, DSS is recommended due to its relative insensitivity to pH variation, unlike TSP[62,63]. However, as mentioned above and in chapter 1, this approach would contaminate the sample, and negative charged DSS at NMR-relevant pHs can interact with the positive charged amino acid side chains (Figure 1.3) and further affect the location of reference value, which would lead to referencing inaccuracies[59]. In addition, the chemical shift of DSS is temperature sensitive and the reference value could deviate due to the different temperatures, if not properly corrected[64].

2.4 Current protein NMR reference detection and correction solutions

Several software packages have been developed and Table 2.1 shows several available programs used by the biomolecular NMR community for correcting referencing in ^{1}H, ^{13}C and ^{15}N chemical shifts[65]. In addition, there are a variety of tools for detecting

protein resonance assignment errors, which can be due to bad referencing. These tools include but are not limited to AVS[66] , PANAV[67], CheckShift[67,68], SHIFTX2[69] and VASCO[70]. Due to the complexity of manual procedures and various experimental factors, approximately 40% of the entries in the Biological Magnetic Resonance Bank (BMRB) have some chemical shift accuracy problems[26,51].

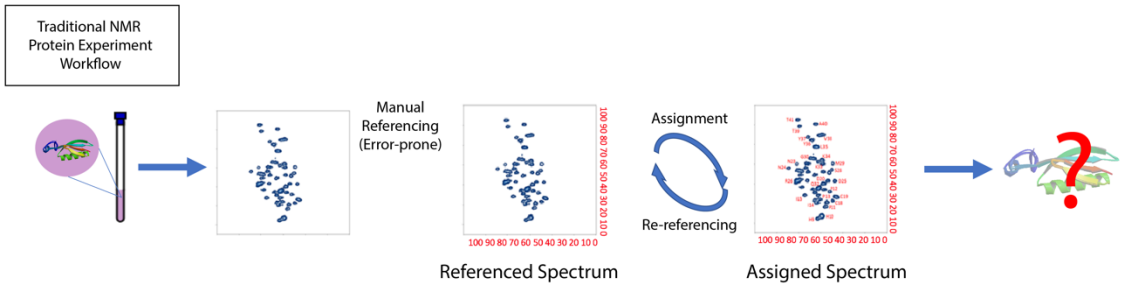

Figure 2.1 Overview of traditional protein NMR referencing workflows.

Unfortunately, current reference correction methods are heavily dependent on the availability of assigned protein chemical shifts or protein structure. One of the best examples is the SHIFTX program [51], which is used by the Re-referenced Protein Chemical shift Database (RefDB) [71] to predict protein ^{1}H, ^{13}C and ^{15}N chemical shifts from the X-ray or NMR coordinate data of previously assigned proteins to check and correct referencing using the companion program SHIFTCOR [71]. Another good example is the linear analysis of chemical shifts (LACS) method, which was developed by the National Magnetic Resonance Facility at Madison and the associated Biological Magnetic Resonance Bank (BMRB) and employs assigned chemical shifts to directly calculate a reference correction[51]. However as shown in Figure 2.1, the traditional workflow requires a manual referencing at step 2 to resolve the assignment initially, by refinement of referencing through a trial and error process. This dependence on assigned chemical shifts

creates a vicious cycle between referencing and assignment in NMR spectra analysis: a correct chemical shift reference is required for good resonance assignment, and a good resonance assignment is needed to validate and correct chemical shift referencing. From a statistical analysis perspective, neither chemical shift referencing nor resonance assignment can be assessed independently of the other.

Table 2.1 Protein chemical shift re-referencing and assignment evaluation software.

Program	Detects or performs shift referencing	Detects gross assignment errors	Distinguishes assignment errors from referencing errors	Requires assigned chemical shifts	Requires 3D structure
CheckShift [60]	Yes	No	No	Yes	No
LACS [51]	Yes	No	No	Yes	No
PANAV [60]	Yes	Yes	Yes	Yes	No
SHIFTX & SHIFTCOR [5]	Yes	Yes	Yes	Yes	Yes
SPARTA+ [21]	Yes	No	No	Yes	Yes
VASCO[70]	Yes	Yes	Yes	Yes	Yes
AVS [51]	No	Yes	No	Yes	No

2.5 Biological context of protein NMR

Stating the obvious, protein NMR research is focused on the characterization of specific proteins. As the main functional unit of all cells and as one of the major products of a gene, proteins are generated mainly from 20 amino acids in a liner fashion (primary) and play many important roles in all levels of biology.

The building elements of protein are amino acids. They are organic compounds contain a carboxy group, an amino group, a hydrogen atom, and a variable side-chain residue (R) as showing in Figure 2.2. Only L-amino acids are found in proteins, and D-amino acids are found in bacterial several walls. Among the common 20 amino acids (see

Figure 2.4.) typically observed in proteins, only glycine is not chiral due to the proton as the R group on the alpha carbon.

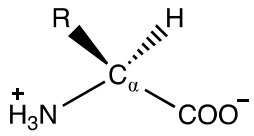

Figure 2.2 Amino acid as a zwitterion in a typical physiological environment.

Under the typical human physiological environment, which roughly ranges between 6.9 to 7.4, amino acids exist as zwitterions, molecules that possess both a positive and a negative charge (Figure 2.2), although, the R group could contains additional acidic or basic group that gives a pK_a value depending on the unique local environment created by the sounding side-chains and this is the very issue mentioned earlier on internal reference standard[72]. In this dissertation, one important amino acid, cysteine, were considered as having two possible chemical states, cysteine and cystine.

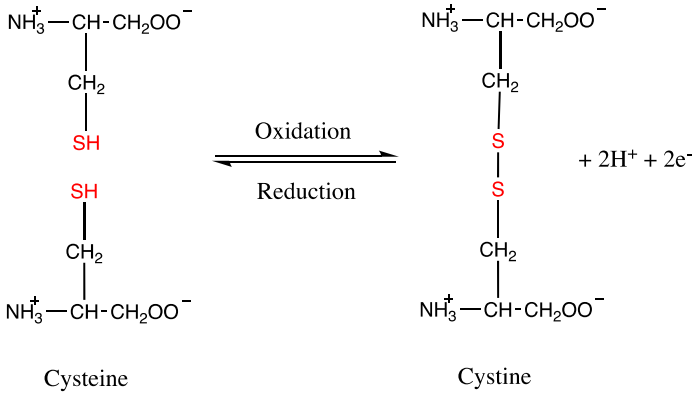

Figure 2.3 Amino acid cysteine's two state. Left: reduced state. Right: oxidized state.

Two cysteines can form a disulfide bond between their thiol groups (-SH) in a process called oxidation, and the resulting residues will be called cystine as shown in Figure 2.3. This disulfide bond, often called a disulfide bridge is commonly observed as a

stabilizing event during protein folding (tertiary structure formation). However, these two states for cysteine yield drastically different chemical shift distributions[68] as shown in Figure 3.7 from Chapter 3. In our protein NMR data analysis, we found it more appropriate to classify cysteine into two separate amino acid states.

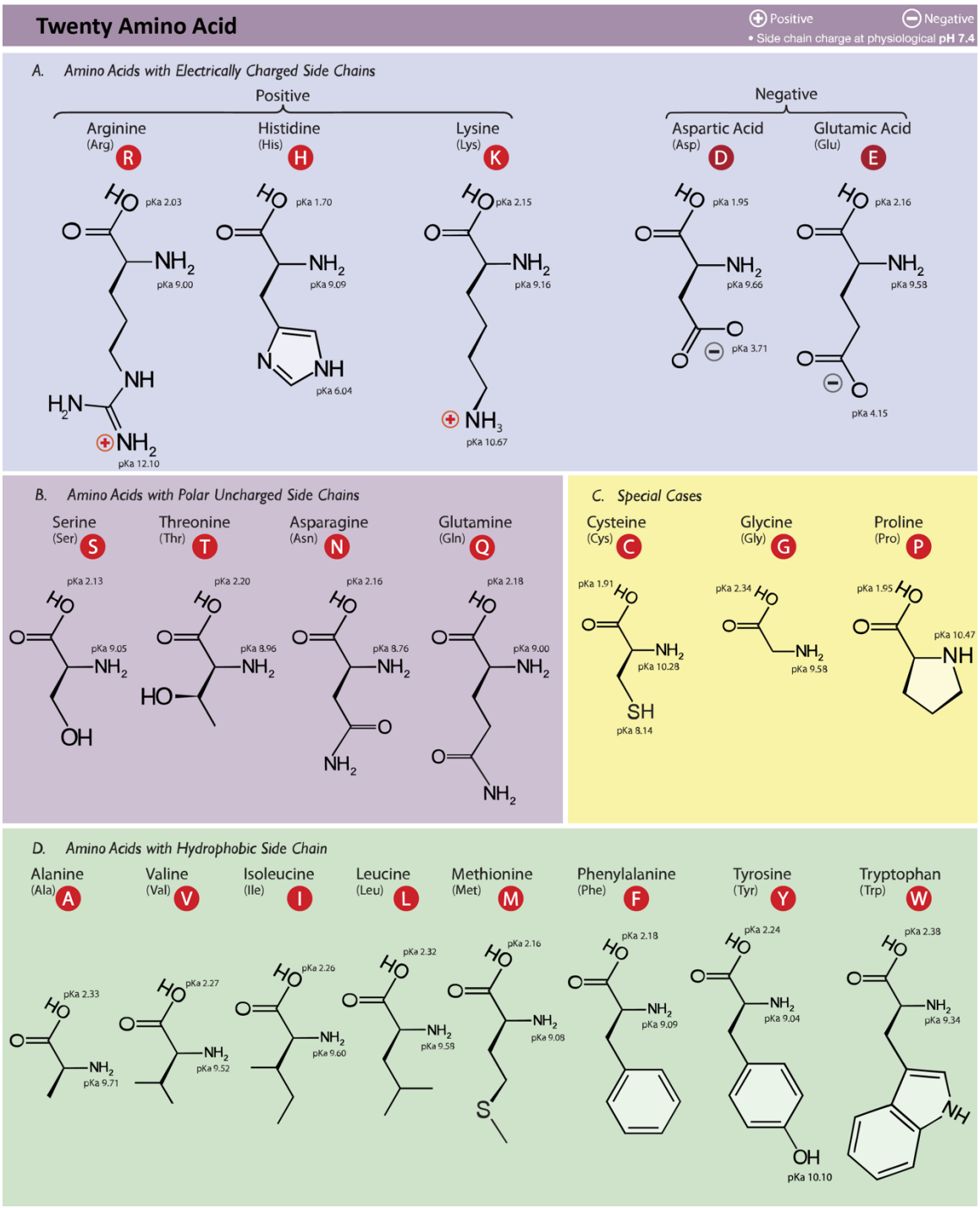

Figure 2.4 20 amino acids and their three- and one-letter conventions[73].

2.5.1 Protein structure and function

As described above, proteins, or peptides, are assembled mostly from the common 20 amino acids. These amino acids forms long linear "chains" by condensation of amino acids through peptide bonds, which the carboxyl group of one amino acid links to the amide group of another amino acid by elimination of water. The partial double-bond characteristics of the peptide bonds, as the lone pairs of electrons on the amine nitrogen are delocalized, which only allows a planar conformation (torsion angles of 0 and 180, cis and trans respectively, and trans conformation is almost universal due to the steric hindrance, except proline.)

This sequence of amino acids is called the primary structure of a protein, and the order of the amino acids in the sequence determines how the protein folds in three dimensions, which ultimately defines the protein functionality[73-80]. The secondary structures present in a protein represent any regular, repetitive folding pattern of the primary structure locally (Figure 2.5). Two major types of secondary structures recognized by the RefDB database, are α-helix and β-strand (Figure 2.6) and any region of the primary structure that cannot fits in these two categories are classified as coils. The major driver of secondary structure is hydrogen bonding between amino acids in a repetitive pattern. An α-helix has 3.6 amino acids per term pattern, and it is stabilized by the hydrogen bonds between the carboxyl group on one amino acid and the hydrogen of the amino group on the next 4^{th} amino acid in the sequence. A β-strand has an extended form in either parallel or antiparallel arrangement, and it is stabilized by the hydrogen bonds on the backbone between hydrogen and oxygen of the peptide bonds of two different strands. Although proline doesn't normally participate in α-helix conformation due to its rigid five-member

ring side-chain which includes the backbone amide nitrogen, it often exists at the beginning of the helix and the turns of between β-strands.

Coils are commonly known as random coils, which is a misunderstanding, since coils are not truly random, instead they have adopted distinctive conformations that are not repeating structures such as an α-helix or β-strand[74]. The random regions of the protein are often highly dynamic or have significant biological functions, and couldn't fit any of the common fixed types of secondary structure. For simplicity, these random regions are also classified into the coil conformation in the protein NMR community.

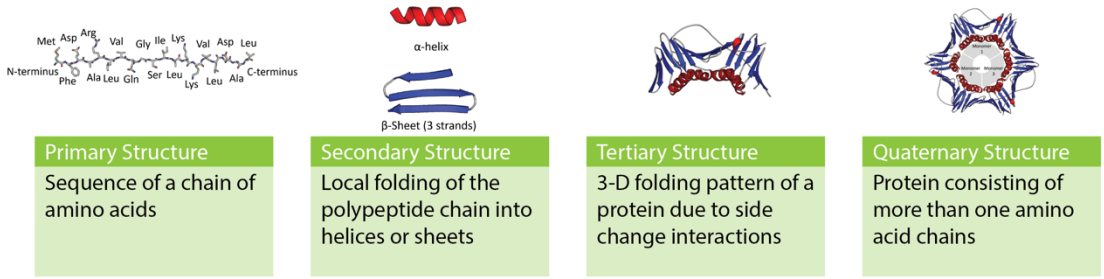

Figure 2.5 Levels of protein structure.

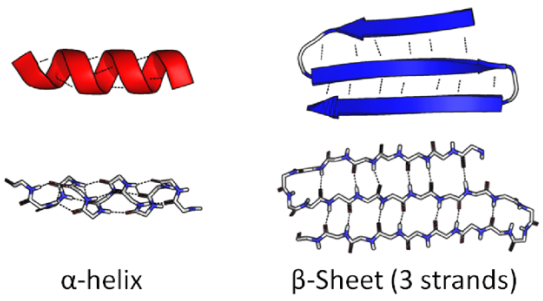

Figure 2.6 Protein secondary structures, left: alpha-helix; right: beta-sheet (strand). (Adapted from image[75].)

Due to (or three if including coil) common types of secondary structure, protein NMM chemical shift data for each amino acid is comprised of multiple unimodal

distributions. Each secondary structure will provide a different electron environment, which leads to a different unimodal distribution. In Chapter 3, I will further describe the statistical significance of these secondary structures from a data analytical perspective.

Tertiary and quaternary (i.e., involving multiple peptide chains) structure of the protein could be considered the global conformation, if secondary structure is the local conformation. In other words, the tertiary structure is from the unique arrangement of secondary structures as shown in Figure 2.5 and Figure 2.6 (adapted from image75.) or often referred to as the protein fold. Many factors contribute to tertiary structure: hydrophobic interactions, dipole interactions, hydrogen bonds, salt bridges, coordination around cofactors, and disulfide bonds. However, detailed understanding of tertiary structure of protein is out of the scope of the dissertation. But generally larger protein fold regions of 100-150 amino acids, commonly associated with a particular function, are known as domains. And protein functionality is determined by the structure of one or more domains.

2.5.2 Protein structure and diseases

From the unveiling of the Omics field through biological and computational analysis, knowledge from genomics, the study of genome (including DNA and RNA), is not adequate for unraveling and characterizing the correlation between gene function and human diseases due to the confounding factors introduced from central dogma such as transcription and post-translational modification (Figure 2.7)[76-78]. Proteins are directly involved by function or malfunction in human diseases, and protein structure fundamentally determines a protein's function. Therefore, protein research have an established long history on the identification of biomarkers for disease screening,

diagnosis, classification and monitoring[79]. In addition, recent literature on potential application of structural proteomics in the field of oncology has demonstrated that protein research, especially protein structure and dynamic play a fundamental role in cancer drug designing, revealing cellular regulatory pathway and personalized therapy for cancer patients[80-83].

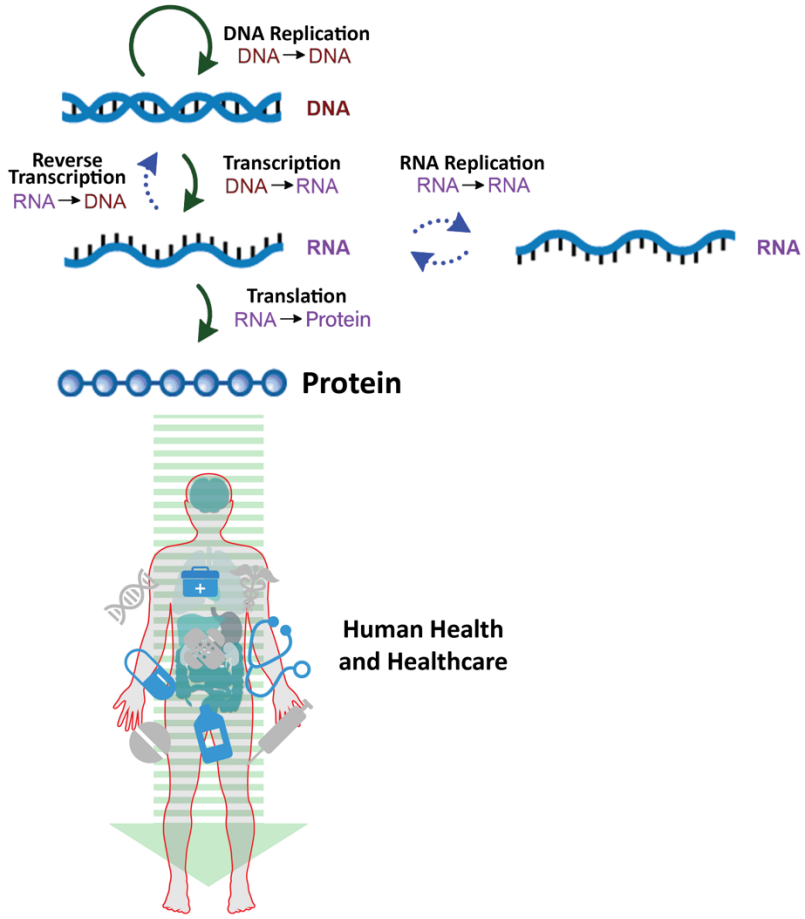

Figure 2.7 Central dogma of molecular biology and human health.

CHAPTER 3. PROTEIN NMR DATA OVERVIEW

3.1 Data source for this dissertation

All of the statistics utilized for this dissertation are from he Re-referenced Protein Chemical shift Database (RefDB)[71]. RefDB is a secondary database derived from a primary database, the Biological Magnetic Resonance Bank (BMRB) [26]. As mentioned in Chapter 2, chemical shifts are relative measurements, which are prone to inconsistencies. Moreover, any inconsistencies in the chemical shift measurements could distort the subtle but rich source of structural and dynamic information present in these measurements. Since 1991, BMRB has served as an archive for interpreted and raw NMR experimental datasets that allow a biomolecular NMR researcher to systematically assemble, compare, and interpret a variety of NMR measurements, especially chemical shifts, across the database. However, these entries are deposited by researchers across the NMR community that utilize a wide range of NMR experiments and data analysis procedures. As a result, roughly 40% of entries in the BMRB contain referencing inconsistencies[26,84]. These data quality issues limit easy reuse of the BMRB, especially global analyses across the BMRB, which became the impetus driving the original development of the RefDB.

The RefDB contains a subset of the BMRB entries that are carefully and properly re-referenced according to the IUPAC/IUB convention[71]. The re-referencing procedure involves using X-ray or NMR coordinate data to estimate protein ^{1}H, ^{13}C and ^{15}N chemical shifts via SHIFTX then compare estimated results with the observed shifts in BMRB via SHIFTCOR[85]. RefDB provides a standard chemical shift resource for the protein NMR community; however, even this resource for protein NMR chemical shifts has some limitations, as is pointed out in Chapter 5.

3.1.1 Normal distribution

The Normal, or Gaussian, distribution is the most important and most widely used distribution. This bell-shape curved distribution can be well approximated in all manner of data that appear to be distributed normally: human height, IQ scores, grades, productions, and chemical shift values are non-exception. Relative frequency distribution can be obtained by suitably normalized frequency distribution, aka. histogram as showing in Figure 3.1. The one-dimensional Normal distribution is determined by just two parameters: mean μ and standard deviation σ of the data. The very definition of Normal distribution $N(\mu, \sigma)$ is: $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$

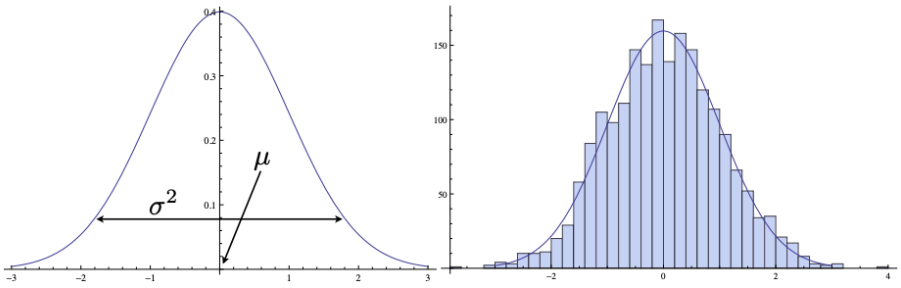

Figure 3.1 Normal distribution plot and histogram.

The mean, μ, controls the location of the peak of the distribution, and σ controls the dispersion of the distribution and the larger the value σ is, the "fatter" the distribution appears. In the scope of the dissertation, the data I used in the BaMORC method are from alpha and beta carbons (Figure 3.2). Thus, a bivariate or 2-D Normal distribution is most appropriate, and a third parameter besides mean and standard deviation is introduced as follows.

The theory of correlation between two variants is an important concept in the mathematical statistics utilized in this dissertation. The bivariate Normal distribution

is an extension of the familiar univariate Normal distribution. Similarly, the probability density function is $p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{z}{2(1-\rho^2)}\right]$

Where $z = \frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}$ and $\rho = cor(x_1, x_2) = \frac{Cov_{1,2}}{\sigma_1\sigma_2}$ is the correlation of x_1 and x_2 and $Cov_{1,2}$ is the covariance. And the covariance is the third parameter for a bivariate distribution, which provides a measurement of strength of the correlation between two random variables. In this context, it is the correlation between chemical shifts of alpha and beta carbons from the same amino acid residue.

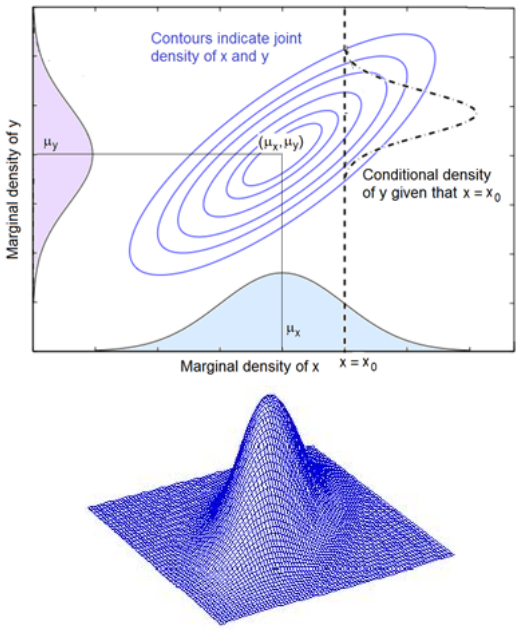

Figure 3.2 Bivariate Normal distribution. Top: 2-D plot of bivariate normal distribution; bottom: 3-D plot of bivariate Normal distribution.

For uncorrelated variates $Cov(x_1, x_2) = 0$; however, if the variables are correlated in some manner, the covariance will be nonzero: if $Cov(x_1, x_2) > 0$, then x_2 tends to increase as x_1 increases, and visa verse. Conventionally, covariance is included into the covariance matrix along with the variance (squared standard deviation) $\Sigma = \begin{bmatrix} \sigma_1^2 & cov \\ cov & \sigma_2^2 \end{bmatrix}$.

The importance of the covariance is due to the correlation between the chemical shifts of the alpha and beta carbons.

3.1.2 The Central Limit Theorem

The name of "the Central Limit Theorem" has many implications; however, the theorem, that most commonly referred to by this name is the following:

The Central Limit Theorem (CLT)

Let $x_1, x_2, \dots, x_n$ be a sequence of random variables that are identically and independently distributed, with mean $\mu = 0$ and variance σ^2. Let $S_n = \frac{1}{\sqrt{n}}(x_1 + x_2 + \cdots + x_n)$. Then the distribution of the normalized sum S_n approaches the Normal distribution of $N(0, \sigma^2)$, as $n \to \infty$.

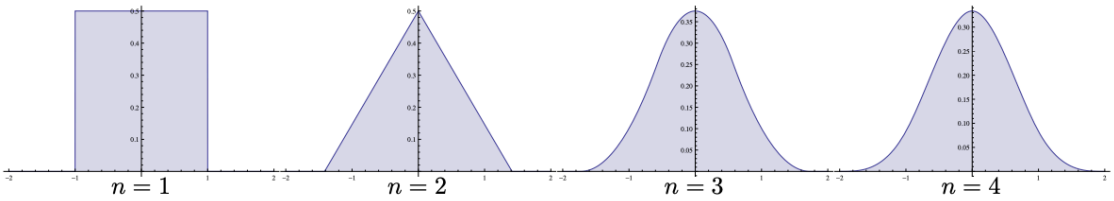

Figure 3.3 A random variable with uniform distribution over [-1,1] added to itself repeatedly. After only four summations, the resulting distribution is very close to a Normal distribution.

Although this seems to be very counterintuitive, the CLT simply states the summing distribution of X will obtain a Normal distribution in the limit, where X can be any distribution with mean of 0 and variance σ^2. Figure 3.3 shows a random variable with uniform distribution over [-1,1] added to itself repeatedly. After only four summations, the resulting distribution is very close to a Normal distribution. The alternative version of the CLT states, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with mean μ and finite variance σ^2, will be approximately

normally distributed with sample mean μ and sample variance $\frac{\sigma^2}{\sqrt{n}}$, regardless of the underlying distribution. And similarly, the CLT applies to bivariate Normal distributions in the same manner.

3.1.3 The Chi-squared distributions.

The following theorem clarifies the relationship between the Normal distribution and the Chi-squared distribution. And the density estimation algorithm from the BaMORC reference correction method uses the chi-squared distribution with two degrees of freedom.

Theorem. If X is normally distributed with mean μ and variance $\sigma^2 > 0$, then: $V = \left(\frac{X-\mu}{\sigma}\right)^2$ is distributed as a chi-squared random variable (χ^2) with one degree of freedom.

And the bivariate version of the theorem is: $(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \sim \chi_2^2$

Proof is showing that the probability density function of the random variable V is the same probability density function of a chi-squared random variable with 1 degree of freedom and the bivariate version can be established in same manner and omitted here, thus we only need to show

$$g(v) = \frac{1}{\Gamma\left(\frac{1}{2}\right)^{\frac{1}{2}}} v^{\frac{1}{2}-1} e^{-\frac{v}{2}}$$

And the cumulative distribution $G(v) = P(V \leq v) = P(Z^2 \leq v) = 1$, where Z follows the standard Normal distribution $N(0, \sigma^2)$ and $V = Z^2$. A proof is out the scope of this dissertation[92]. The chi-squared distribution with two degree of freedom is showing in Figure 3.4

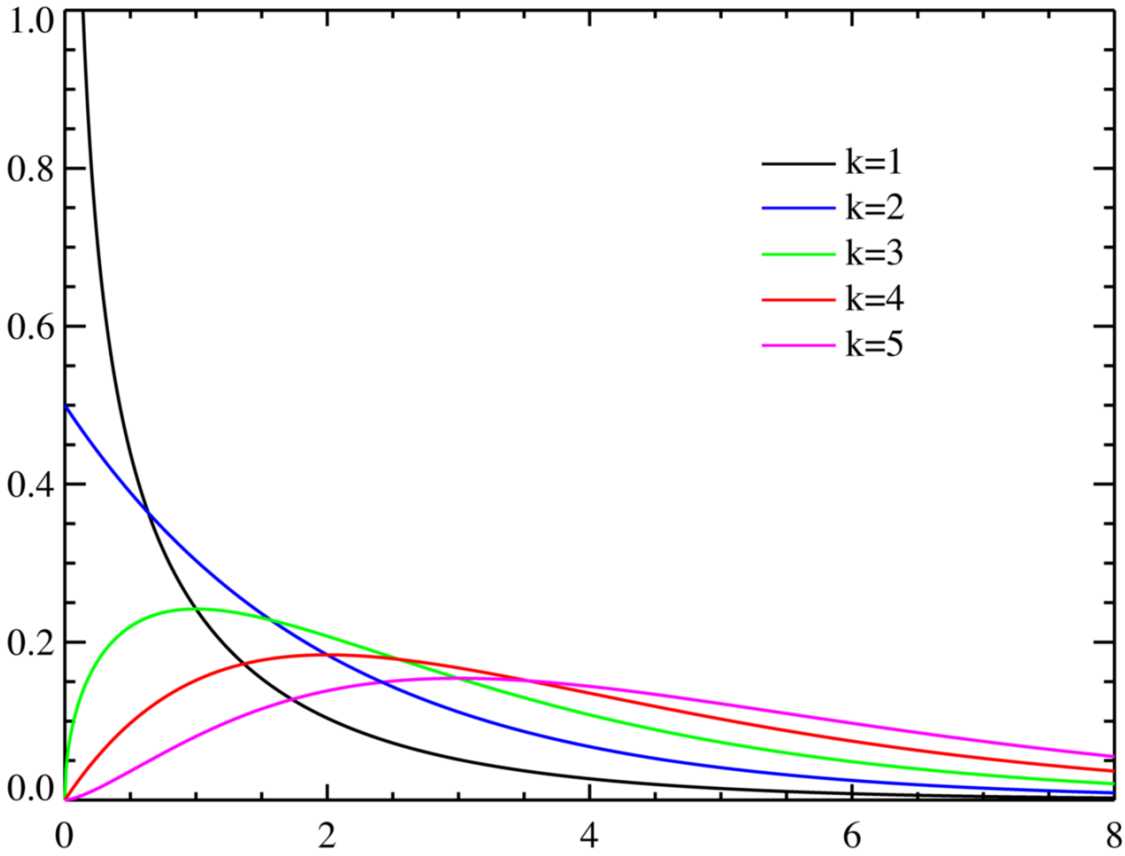

Figure 3.4 Chi-squared density distribution with k=1, 2, 3, 4, 5 degrees of freedom (Figure adapted from WikiMedia.).

3.2 Protein NMR carbon (alpha and beta) data analysis

For the scope of this dissertation, we are only concerned with the alpha carbon and beta carbon chemical shifts. We downloaded the 2162 available protein chemical shift datasets from the Re-referenced Protein Chemical shift Database (RefDB)[1] on May 4th, 2015 [86]. The developers of the RefDB have carefully corrected the referencing of ^{1}H, ^{13}C, and ^{15}N chemical shifts in BioMagResBank (BMRB) entries using the SHIFTX-predicted chemical shifts based on corresponding 3D protein structures in the Protein Data Bank (PDB), which is managed by the international collaboration known as the worldwide Protein Data Bank (wwPDB) (Berman et al. 2007). Among the 2162 RefDB entries, we employed 1557 that contained both C_α and C_β chemical shifts, both to derive the necessary

statistics and then to subsequently test our methods. Secondary structure specific information was likewise downloaded and extracted from the RefDB website.

3.2.1 Protein NMR carbon chemical shifts distribution

For each RefDB entry, we first parse the text data files with the extension of ".str.corr", which are mostly in version 2 of the NMR-STAR format[26], with additional sections added by RefDB, with a short R script that uses crafted regular expressions to clean and convert the relevant assigned chemical shift data into a tab-based format for easier parsing later. The reason for this conversion step is to remove unnecessary metadata, missing values, blank spaces, and section breaks. In this conversion, we retained the full sequence, residue position, amino acid typing, secondary structure, and C_α and C_β chemical shift information. Statistics such as the mean and standard deviation were also calculated from the resulting data and verified using the results reported on the RefDB website. Based on amino acid and secondary structure, we subdivided the data into 60 classes based on 20 amino acid types and 3 secondary structure types. In the early part of the methods development, we ignored the glycine classes and only employed the other classes representing the 19 amino acids with C_β resonances.

We extracted all relevant ^{13}C chemical shift entries (datasets) from the processed data as mentioned in the previous paragraph. Each dataset contains the protein sequence and the corresponding NMR chemical shifts. One point worth mentioning is that most of the datasets are not complete: i.e., there are fewer assigned residues than would be expected from the protein sequence. However, missing resonance assignments are common due to a myriad of experimental conditions, especially conformational flexibility in the protein structure that leads to intermediate chemical exchange. Chemical exchange

occurs from conformational change or binding events between chemically distinct environments on microsecond to millisecond time scales that lead to an averaging of the chemical shifts of nuclei in each distinct chemical environment. Fast exchange leads to a weighted average chemical shift while slow exchange allows for the detection of multiple chemical shifts representing each distinct chemical environment; however, intermediate exchange can lead to a null event due to line broadening with no detectable peak[87]. Using the secondary structure information accompanying the NMR chemical shift data provided by the RefDB, we associated residue-specific C_α and C_β chemical shifts and then sub-grouped them by amino acid and secondary structure type, as showing in Figure 3.5 for 19 of the 20 common amino acids (not including glycine) in proteins and for the secondary structure types helix, sheet, and coil.

For all of the amino acids, the univariate C_α and C_β chemical shift distributions are multimodal, with most of the modes being secondary structure specific [6,7]. One important assumption in this project is that the separate C_α and C_β chemical shift modes follow a chi squared distribution with two degrees of freedom, also expecting that the separate C_α and C_β distributions each follow a Normal distribution. Goodness-of-fit tests for a Normal distribution (Table 3.1) do indicate that these chemical shifts are roughly normally distributed, i.e. sample data reasonably fits a Normal distribution when tested with expected sample sizes[88-92].

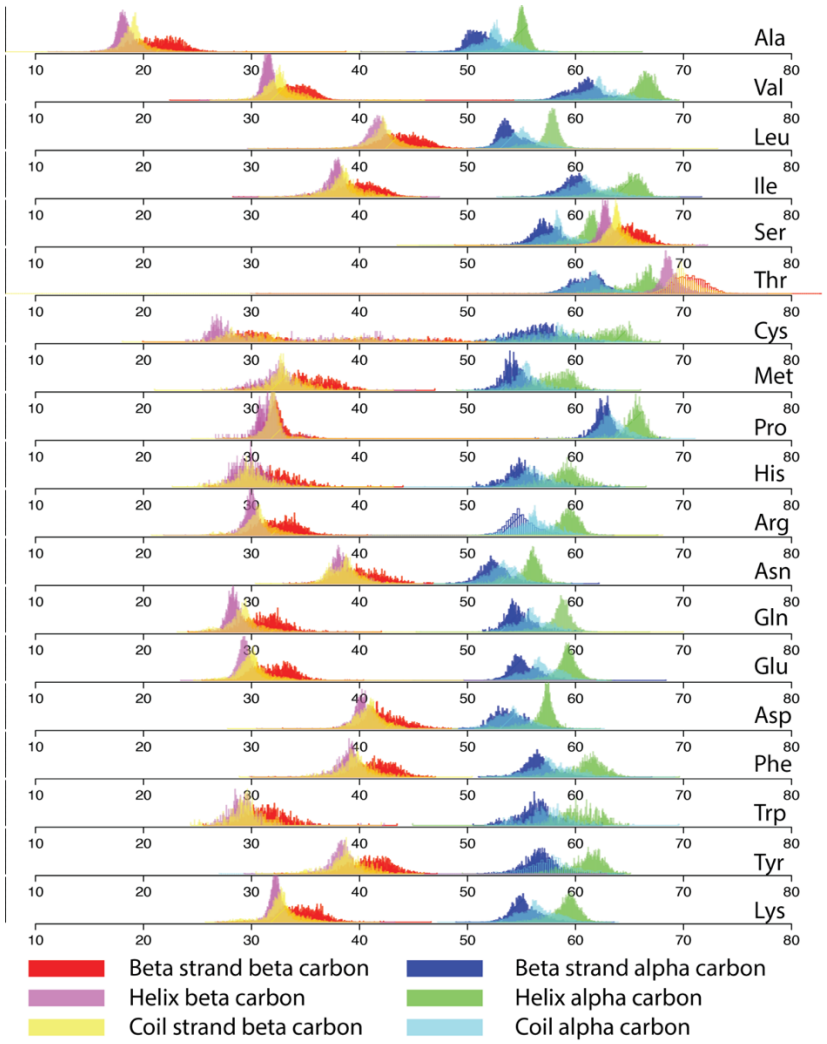

Figure 3.5 Univariate chemical shifts distribution of alpha and beta carbon from RefDB. Please note that the distribution plot here are in reverse of how a spectroscopist typically views chemical shift spectral data.

Table 3.1 Goodness-of-fit tests for Normal distribution.

	Anderson-Darling Test[89]	Lilliefors Test[90]	Pearson Chi-square Test[91]	Shapiro-Francia Test[92]
Alanine	37.95%	44.58%	55.15%	39.33%
Cystine	26.82%	25.95%	28.47%	30.60%
Aspartate	41.82%	49.88%	60.82%	38.67%
Glutamate	30.72%	36.65%	52.00%	30.50%
Phenylalanine	40.78%	48.90%	67.28%	39.68%
Glycine	43.57%	51.03%	71.80%	42.57%
Histidine	55.38%	63.73%	77.27%	50.75%
Isoleucine	41.78%	47.07%	57.18%	40.25%
Lysine	32.33%	38.03%	51.82%	29.12%
Leucine	37.05%	43.55%	54.60%	34.88%
Methionine	36.20%	45.62%	60.05%	33.48%
Asparagine	44.82%	53.37%	64.22%	39.05%
Proline	21.48%	32.48%	46.48%	16.10%
Glutamine	29.67%	35.23%	48.42%	29.35%
Arginine	35.42%	41.05%	56.42%	36.03%
Serine	26.87%	31.78%	43.43%	25.05%
Threonine	30.67%	35.97%	44.78%	27.07%
Valine	40.65%	46.70%	55.50%	40.45%
Tryptophan	54.57%	65.57%	74.47%	50.08%
Tyrosine	47.10%	54.33%	68.17%	44.45%

The one-dimension distribution couldn't fully capture the characteristics of a bivariate distribution such as the correlation between the alpha and beta carbon chemical shifts. Also, we have a general understanding of the key statistical features of the chemical shifts such as chemical shifts statistics generally indicate that the alpha carbon (C_α) is around 50-70 ppm and the beta carbon (C_β) is around 15-45 ppm, with exceptions for glycine, threonine, and serine, due to the lack of a side chain for glycine and a bound oxygen atom to C_β for serine and threonine. Also, there is a relationship between the secondary structure and chemical shifts. The deviation from random-coil chemical shift is referred to as secondary shift, which is denoted as δ for residue k. The trend between

chemical shifts and secondary structure in proteins led to the definition of "secondary structure shifts" or simply "secondary chemical shifts. The secondary chemical shift $\Delta\delta_S{}^i$ of a particular protein nucleus i is defined as: $\Delta\delta_S^i = \delta_{obs}^i - \delta_{r.c}^i$, where δ_{obs}^i is the observed chemical shift and $\delta_{r.c}^i$ is the corresponding random coil value as shown in the Figure 3.6 Histograms of secondary shift distribution in α-helix and β-strand [51].

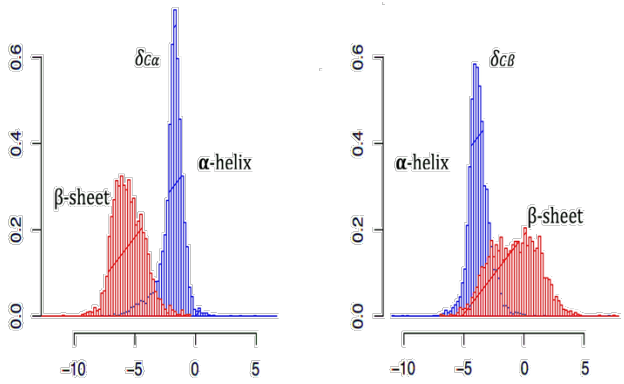

Figure 3.6 Histograms of secondary shift distribution in α-helix and β-strand. The red color represents the β-strand secondary shift distribution and the blue color represents the α-helix secondary shift distribution.

This secondary shift suggested that there is a fundamental correlation between alpha carbon and beta carbon and appears the opposite in α-helix and β-strand secondary structures. Instead of using a univariate distribution approach, we further included the covariance in the bivariate distribution approach as showing in Figure 3.7. Next, we calculated the mean and standard deviation specific to the amino acid and secondary structure type and verified these statistics with the values provided by the RefDB. We then calculated the covariances between alpha and beta carbons. Figure 3.7 illustrates the overlapping alpha and beta carbon distributions for the 20 common amino acids minus glycine, and it demonstrates the reason why simple statistical models are inadequate without considering secondary structure, reduced/oxidized cysteines, and covariances. Figure 3.7a shows the distribution of all the RefDB data with contouring for the 19

common amino acids with both C_α and C_β. Figure 3.7b shows these distributions represented with simple, independent bivariate models for each amino acid, as illustrated by ellipses centered on C_α and C_β chemical shift means, with the axes representing 2 standard deviations and providing approximately 95% coverage of the data. Figure 3.7c illustrates the same independent bivariate models, but with oxidized and reduced cysteines modeled separately. Figure 3.7d illustrates bivariate models with covariance. Figure 3.7e illustrates 60 bivariate models with covariance for the 19 common amino acids, sub-divided by secondary structure categories helix, sheet, and coil and with cysteine further divided into oxidized and reduced forms. These final 60 bivariate models match the observed distributions derived from RefDB data asymptotically and represent a key ingredient in the BaMORC methodology. The alpha and beta ^{13}C chemical shift statistics used in these models are summarized in Table 3.2**.**

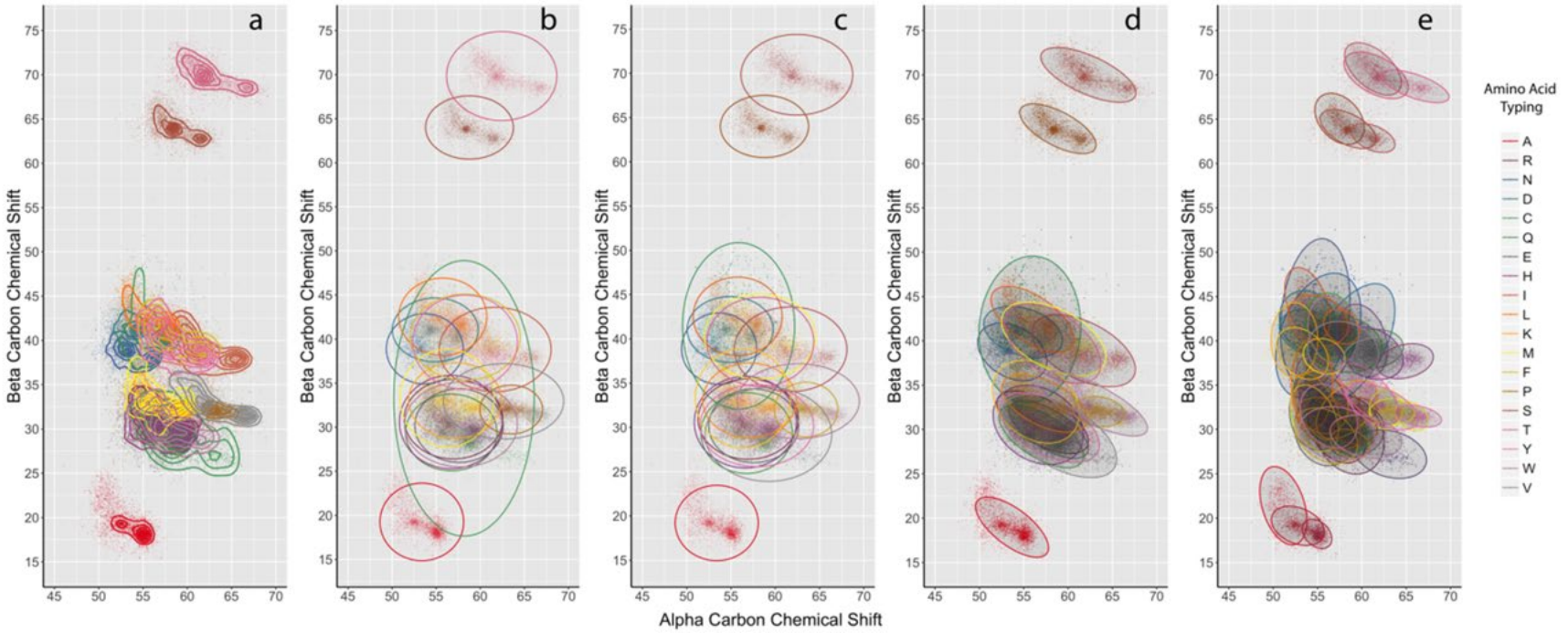

Figure 3.7 2D Distributions of alpha and beta carbon chemical shifts specific to amino acid and secondary structure types. a: the actual distribution of 19 amino acids (excludes glycine due to lack of beta carbon); b: using simple statistics (without covariance) could not model the distributions well, with many overlapping ovals; c: treating cysteine as two distributions achieved a better modeling (without covariance); d: including the covariances further improved the models, allowing a better classification; e: including secondary structure refines the models further.

In Figure 3.7, graph **a** contains the actual, i.e., true, bivariate distributions with density. Graph **b** has statistically modeled distributions without covariance. Graph **c** is the same as **b,** but with cysteines represented as separate distributions. Graph **d** has statistically modeled distributions with covariance. Graph **e** has statistically modeled distributions with covariance for three secondary structure types. And all of the individual 2D distribution of 19 amino acids are shown in Figure 3.8.

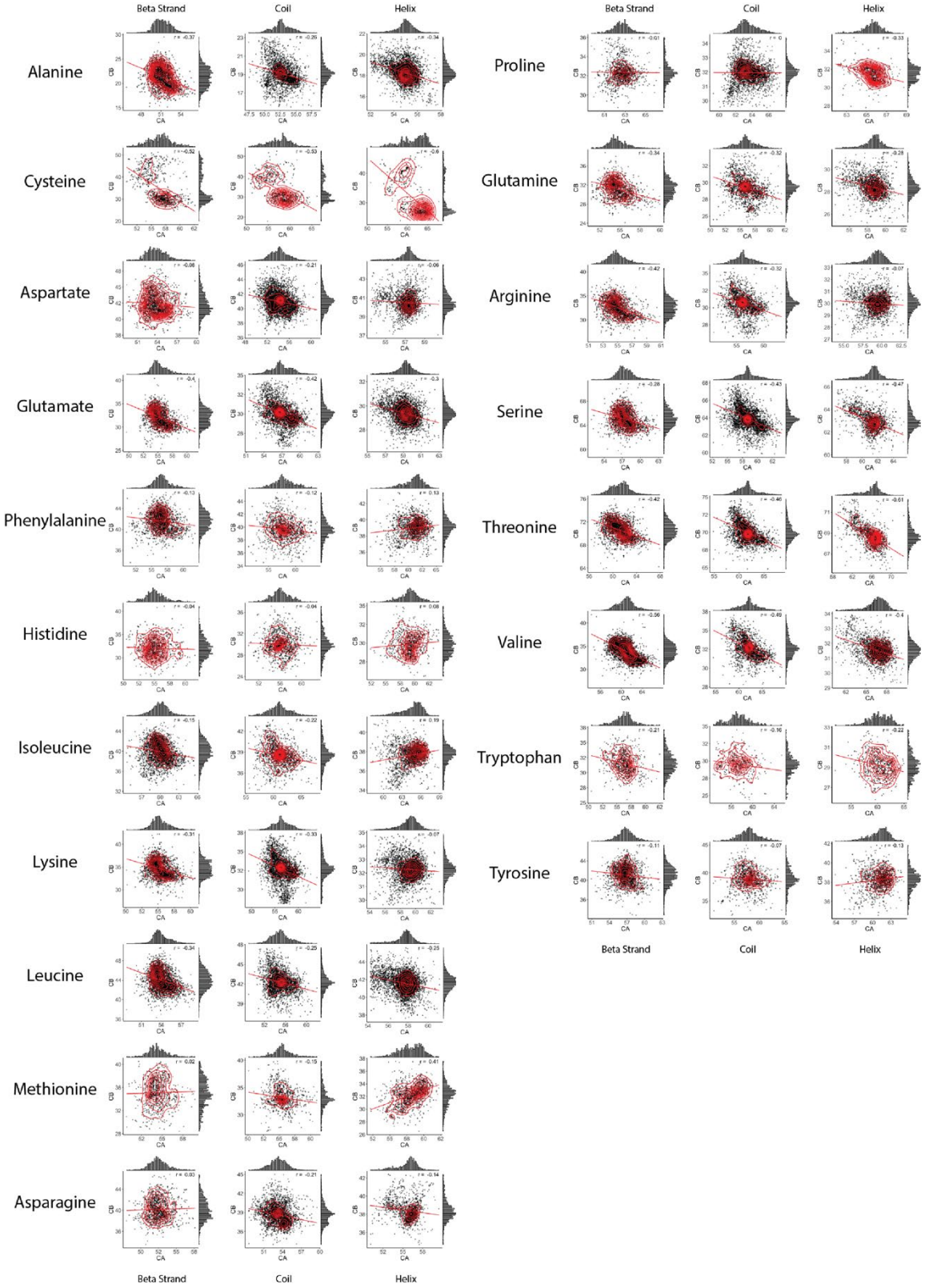

Figure 3.8 Individual 2D distributions for all 19 amino acids.

Table 3.2 The summary of alpha and beta 13C chemical shift statistics used in the statistical models. AA: amino acid name, B: beta strand, H: alpha helix, C: coil.

	Ca Mean			Cb Mean			Ca SD			Cb SD			Covariance		
AA	C	H	B	C	H	B	C	H	B	C	H	B	C	H	B
A	52.84	54.83	51.53	19.06	18.26	21.14	1.64	1.05	1.48	1.26	0.88	2.05	-0.58	-0.31	-0.99
C_o	57.04	59.63	56.02	40.58	39.34	42.98	2.33	2.43	1.72	2.99	2.79	3.88	1.09	1.99	1.05
C_r	57.51	61.58	56.57	29.50	27.47	30.08	2.49	2.89	1.76	1.97	1.37	1.69	-0.37	-0.40	-0.51
D	54.18	56.70	53.87	40.85	40.51	42.30	1.60	1.61	1.64	1.32	1.33	1.62	-0.48	0.05	-0.10
E	56.87	59.11	55.50	30.20	29.37	32.01	1.82	1.16	1.67	1.55	0.99	1.98	-1.00	-0.18	-1.04
F	57.98	60.81	56.65	39.45	38.78	41.54	2.02	1.90	1.59	1.98	1.31	1.74	-0.22	0.32	-0.45
H	55.86	59.04	55.09	29.97	29.54	31.85	1.96	1.74	1.78	2.42	1.46	2.22	-0.17	0.28	0.19
I	61.03	64.57	60.05	38.65	37.60	39.86	1.90	1.74	1.57	1.69	1.15	1.98	-0.72	0.44	-0.44
K	56.59	58.93	55.40	32.79	32.27	34.63	1.78	1.44	1.34	1.67	0.88	1.78	-0.82	0.02	-0.72
L	54.92	57.52	54.00	42.38	41.65	43.79	1.70	1.23	1.31	1.64	1.05	2.00	-0.54	-0.31	-0.80
M	55.67	58.09	54.58	33.36	32.27	35.05	1.54	1.81	1.24	2.26	1.66	2.29	-0.75	1.13	0.10
N	53.23	55.45	52.74	38.55	38.61	40.12	1.51	1.42	1.47	1.41	1.31	2.07	-0.46	-0.20	0.23
P	63.47	65.49	62.64	31.94	31.46	32.27	1.26	1.08	1.03	0.95	0.95	1.20	-0.05	-0.20	-0.02
Q	56.12	58.47	54.83	29.14	28.51	31.28	1.72	1.19	1.41	1.69	0.92	1.93	-0.93	-0.20	-0.84
R	56.42	58.93	55.14	30.66	30.14	32.19	1.94	1.55	1.64	1.67	1.14	1.80	-0.73	0.00	-1.05
S	58.38	60.88	57.54	64.03	63.08	65.16	1.69	1.61	1.40	1.27	1.12	1.51	-0.74	-0.36	-0.52
T	61.64	65.61	61.06	70.12	68.88	70.75	2.07	2.39	1.59	1.33	1.17	1.51	-1.37	-1.37	-0.92
V	62.06	66.16	60.83	32.71	31.49	33.91	2.16	1.55	1.64	1.37	0.72	1.61	-1.33	-0.33	-1.49
W	57.78	60.01	56.41	29.67	29.30	31.50	1.71	1.77	1.87	1.74	1.40	1.70	-0.81	-0.50	-0.64
Y	57.97	60.98	56.83	38.95	38.25	40.97	2.17	1.76	1.71	1.84	1.11	1.85	-0.12	0.20	-0.35

3.2.2 Separating bivariate distributions of alpha and beta carbons for oxidized and reduced cysteine residues

The amino acid cysteine has historically caused substantial inaccuracy in the prediction of amino acid types. Figure 3.5 shows the wide spread of C_α and C_β chemical shifts for the cysteine residue distributions over almost the whole expected C chemical

shift range for the common amino acids. In contrast, alanine exhibits tight, well-behaved, unimodal bivariate distributions for each secondary structure type. The problem of modeling the cysteine distribution as a whole is illustrated by a large bivariate ellipsoid model in Figure 3.7b. The broad cysteine residue distribution hinders the use of expected chemical shift values and variances in calculating the probabilities of amino acid types[51]. The wide cysteine distribution occurs because of the existence of two common side-chain oxidation states for cysteine residues within proteins: the oxidized disulfide-bonded cysteine form and the reduced cysteine form [93,94]. However, while the univariate distributions of individual carbon chemical shifts are broad and indistinct, as shown in Figure 3.5, the cysteine bivariate chemical shift distributions exhibit distinct modes that are specific to different oxidation states and secondary structure types, as illustrated by multiple contoured density centers in the top graphs of Figure 3.9. In contrast, alanine mainly exhibits a single contoured density center for each secondary structure type, as shown in the bottom graphs of Figure 3.9. As the calculated C_α and C_β chemical shift covariances span these extra modes, ignoring them will reduce the amino acid prediction power of the statistical methods utilized in BaMORC. Since the RefDB entries do not indicate the oxidation state of the cysteine residues, we used a K-means clustering method, as described in the Methods, to separate the cysteine residues into two oxidation groups for each secondary structure type, as shown in Figure 3.9. We also employed the convention that the C_o refers to the oxidized form of cysteine while the C_r refers to the reduced form.

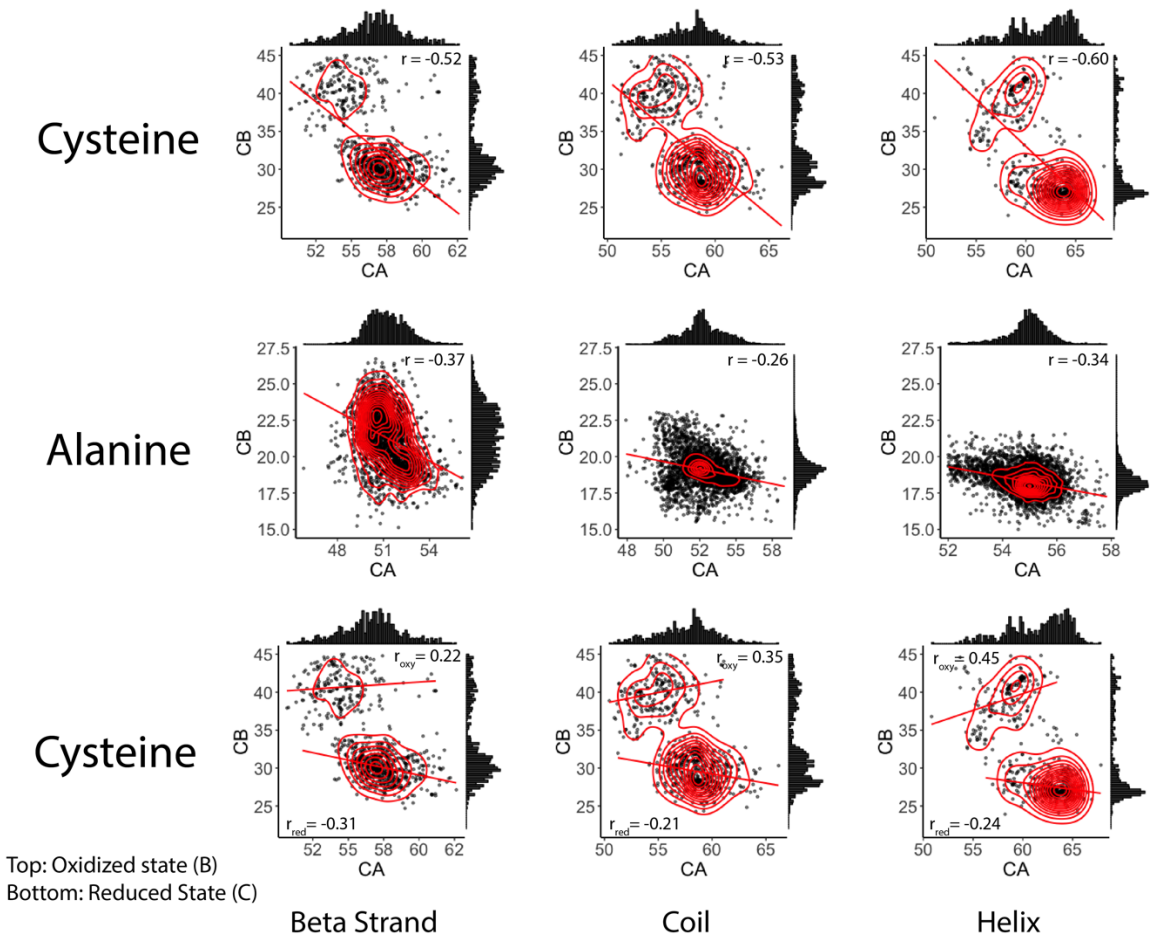

Figure 3.9 Top two panels: Amino acid distributions for alanine and cysteine, with corresponding correlation values. Top: cysteine distributions for each secondary structure were treated as a single distribution, which is obviously inappropriate. Middle: alanine distributions across three secondary structures, which is indeed a single distribution. Bottom: cysteine distributions were treated as two separate bivariate distribution basing on the oxidation state, which is appropriate and gives different correlation values (red lines in the figures represents the regression lines associated with the correlation values).

At the top of the Figure 3.9, the cysteine chemical shifts values were plotted as a single population per secondary structure, which is not convincing due to the two obvious clusters/distributions. At the bottom, we used K-mean cluster algorithm and grouped the cysteine chemical shifts values in two clusters, two separate population based on the oxidation state, with corresponding correlation values, the plot became more appropriate

and visually intuitive. For these two oxidized and reduced cysteines populations, the derived correlation values are shown in Figure 3.9.

3.2.3 K-means clustering of oxidized and reduced cysteine alpha and beta carbon chemical shifts

From Figure 3.9, we concluded that cysteine chemical shifts are too broad and needed to be treated as two different populations based on two oxidation states, reduced and oxidized. It is worth mentioning that even though the statistics from RefDB included two-state cysteines, there are no labels on any specific cysteine in the RefDB NMR data. Therefore, we had no choice but to perform a two-group clustering to separate oxidized and reduced cysteine C_α-C_β pairs. For this purpose, we utilized the K-means clustering machine-learning algorithm[76]. This algorithm requires the expected number of clusters, K, which was two in this specific application. The algorithm begins by selecting K=2 data points as "centroids" and groups each C_α-C_β pair into two clusters based on the smallest Euclidean distance from cluster centroids. Then, it uses iterative techniques to re-calculate the centroids and re-group the data until the centroids converge. To verify the clustering results, we compared the means and standard deviations of the two new subgroups with statistics reported in the RefDB.

3.2.4 Calculating and refining alpha and beta carbon covariances

After grouping all of the RefDB datasets based on amino acid and secondary structure, we calculated the covariance between C_α and C_β for each group. We first calculated the mean (μ) and standard deviation (sd) for C_α and C_β of each group i, as show

in Equation $\mu_\alpha = \frac{\sum_{i=1}^{n} C_{\alpha,i}}{n-1}$; $\mu_\beta = \frac{\sum_{i=1}^{n} C_{\beta,i}}{n-1}$ and $sd_\alpha = \sqrt{\frac{\sum_{i=1}^{n}(C_{\alpha,i}-\mu_\alpha)}{n-1}}$; $sd_\beta = \sqrt{\frac{\sum_{i=1}^{n}(C_{\beta,i}-\mu_\beta)}{n-1}}$.

Then, we used $Cov_{\alpha,\beta} = \frac{\sum_{i=1}^{n}(C_{\alpha,i}-\mu_\alpha)(C_{\beta,i}-\mu_\beta)}{n-1}$ to calculate the covariance $Cov_{\alpha,\beta}$.

The covariance matrix was constructed using $\Sigma = \begin{bmatrix} sd_\alpha^2 & Cov_{\alpha,\beta} \\ Cov_{\alpha,\beta} & sd_\beta^2 \end{bmatrix}$ equation and the matrix representation was employed in the algorithm.

Due to the variation in the quality of the data, the covariances calculated from all of the RefDB data are not representative, causing the reference correction values to be less accurate. When C_α and C_β chemical shift data are collected from two separate NMR experiments, two independent samples of chemical shifts are generated. Similar to the batch effects, these two samples are independent and the correlation between the α and β carbons are weakened or even destroyed. Thus, it was necessary to select a subgroup of data and re-calibrate the covariance. The data filtration procedure is shown in Figure 3.11.

We employed the root mean squared deviation (RMSD) as the criterion for selecting subgroups. The RMSD is recorded in every data file in the RefDB. The RMSD is a measurement of the confidence interval of the population mean (mean of the difference between the calculated and observed shifts) for each single data point. This statistic is calculated from Student's t-test. The higher the RMSD value, the less accurate the corrected data. In our methodology, we have two RMSDs from the C_α and C_β nuclei. To select the best datasets, we need lower individual RMSDs, a smaller difference between the two RMSDs, and, simultaneously, the maximum difference in the correlation between two subgroups (useful data and non-useful data). Thus, we first compared the two RMSD values, using the RMSD comparison equation Q, as shown in Figure 3.11. The rationale behind this transformation is the minimization of the difference between RMSDs, which

is the absolute difference in the numerator under the cubic root, and the minimization of individual RMSD values by dividing the numerator by the sum of their absolute values. In this context, the cube root is a standard statistical transformation method, allowing a very skewed distribution to approximate a Normal distribution[95,96] , as shown in Figure 3.11. Then, we divided the data into two groups based on the cutoff point from the Q values, calculated the correlations r_1 and r_2 of both groups, and then used the correlation test to calculate the p-value, as shown in steps 2 and 3 of Figure 3.11. By recursively applying steps 2 and 3, we identified the smallest p-value as the final cutoff point. All of the data (per-structure) that provide Q values smaller than the cutoff point is included in the datasets to further refine the covariance.

3.2.5 Refining alpha and beta carbon covariances

The re-referenced C_α and C_β chemical shifts in the RefDB are derived from BMRB entries that are based on protein resonance assignments derived from multiple NMR spectra. Unfortunately, it is unclear from a BMRB entry whether a given set of alpha and beta ^{13}C chemical shifts are derived from the same NMR spectrum or from multiple spectra, except when assigned peak lists are included, which is the case for only a small fraction of BMRB entries. The C_α and C_β chemical shifts from different spectra can be misregistered (i.e. shifted out of register with each other), weakening the covariance calculated between these chemical shifts. For instance, C_α and C_β chemical shifts could either be from the same experiment, for instance an HNcoCACB NMR experiment or two experiments, for instance HNcoCACB and HNcoCA NMR experiments (Figure 3.10). If C_α and C_β chemical shifts are reported from two separate experiments, the covariance or

joint variability can be lost, destroying the ability to accurately calculate the covariance from a dataset. Just as the requirement for many biological measurements, the chemical shifts for both alpha and beta carbons should be measured from the same experiment, i.e. measurable phenomenon.

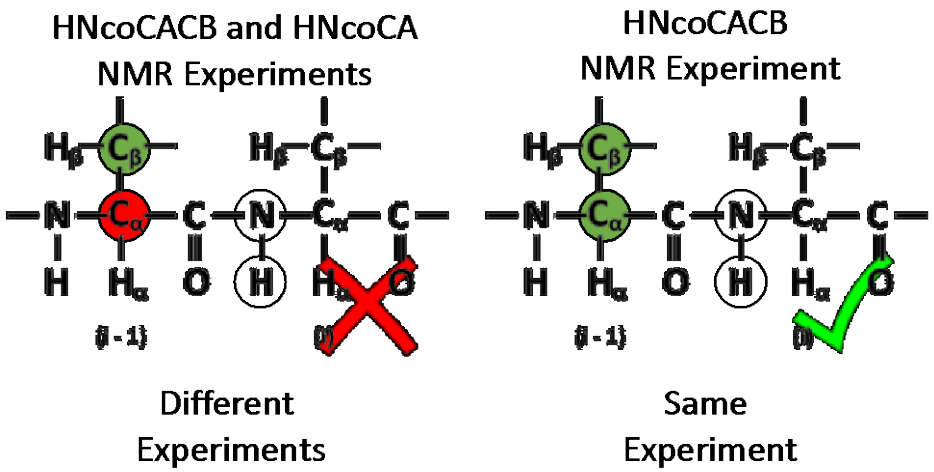

Figure 3.10 Comparison of two sources of RefDB chemical shifts for alpha and beta carbon. Right: alpha carbon chemical shifts are from an HNcoCA experiment and beta carbon chemical shifts are from an HNcoCACB experiment. Left: both chemical shifts are derived from the same HNcoCACB experiment.

Therefore, we used quality control measures provided by the RefDB to evaluate the performance of the RefDB referencing correction and used this to create a criterion for selecting a subset of entries for deriving amino acid- and secondary structure-specific covariances between C_α and C_β chemical shifts.

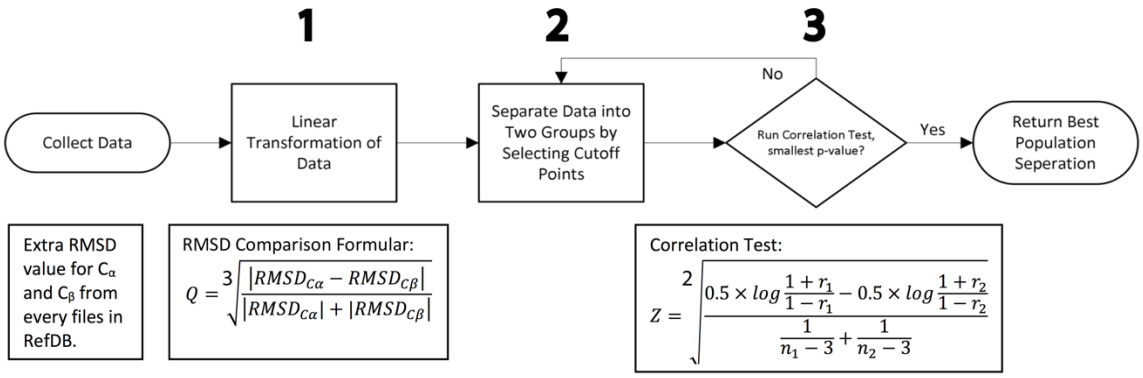

Figure 3.11 Data selection algorithm for re-calculating covariances.

Specifically, we employed the absolute difference between alpha and beta carbon root mean squared deviations (RMSD) from SHIFTX2-predicted and observed chemical shifts to order entries as shown in Figure 3.11. Based on the RMSD values provided with RefDB datasets, we (1) performed a cubic root transformation; then (2) separated the datasets into two groups based on the Q values and a small p-value against the other subgroup. We then repeated steps (2) and (3) to identify the subgroup with the best sample for covariance calculations. Next, we incorporated entries in a best-first manner into the calculation of C_α and C_β chemical shift correlations until the sum of the absolute value of these correlations were maximized. After maximization, 729 of the 1557 entries from the RefDB were selected to calculate covariances. The entire workflow is detailed in the next chapter. In addition, Figure 3.12 shows the differences between the covariances calculated before and after optimization.

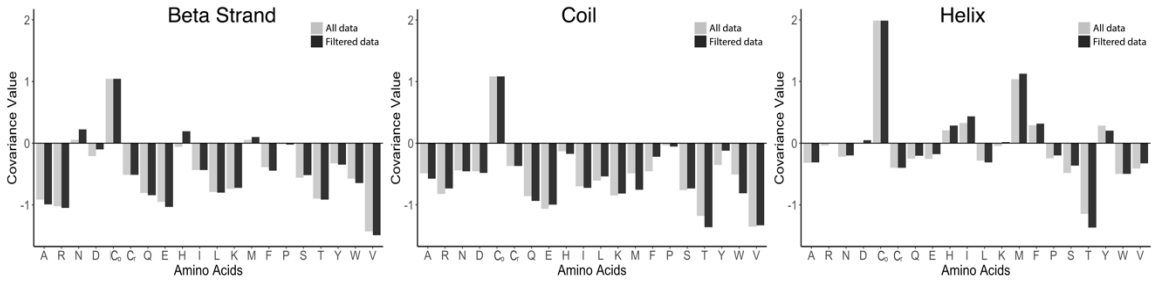

Figure 3.12 Comparison of covariance values calculated using all of the data from RefDB or using filtered data only. Almost all the covariances has a certain level of difference, though bigger covariance value does not suggest a better approximation of the true covariance statistics, and some even have a sign change, i.e. from positive to negative or negative to positive.

For all three secondary structures, most of the covariances increase in magnitude. Some of the covariances even show a sign change, which provides a significant improvement in prediction outcomes. Note: the C_o stands for the oxidized cysteine state and C_r for the reduced cysteine state.

CHAPTER 4. PROJECT DESIGN OVERVIEW

4.1 Introduction

This chapter provides a high-level overview of the algorithm and data structures necessary to model and solve the protein NMR reference correction problem. By now, I hope I have already convinced you that poor chemical shift referencing, especially for ^{13}C in protein Nuclear Magnetic Resonance (NMR) experiments, fundamentally limits and even prevents effective study of biomacromolecules via NMR, including protein structure determination and analysis of protein dynamics. To solve this problem, we constructed a Bayesian probabilistic framework that circumvents the limitations of previous reference correction methods that required protein resonance assignment and/or three-dimensional protein structure as shown in Figure 4.1. The traditional workflow requires a manual referencing at step 2 to resolve the assignment initially, followed by refinement of referencing through a trial and error process.

4.2 Rationale for using RefDB and its limitation

In this statistical model building and data analysis methods development, we utilized RefDB data for several pragmatic reasons. First, the RefDB is the best-referenced large carbon chemical shift dataset that is currently available. Second, we can treat RefDB as a gold standard for evaluation purposes, because it represents a systematic reference correction subset of the BMRB and was the only large dataset we could reasonably use for evaluation of performance. Third, we chose real datasets over simulated datasets, because of the difficulty in generating simulated datasets that represented the complexity of real datasets adequately enough to evaluate performance [28]. Simply stated, there was too high

a possibility of overestimating performance with simulated datasets that inadequately reflected the complex deviations in carbon chemical shifts of real datasets.

However, it is well recognized in the field that deposited NMR chemical shift data have inaccuracies, and that the RefDB still include errors. Because of these errors, the statistics that we extracted from the RefDB data might not be representative of protein NMR as a whole. Although, a number of algorithms and methods attempt to correct the reference, most of these approaches rely on the assignment of the sequence at the end of the data analysis stage. Our algorithm was built using derived statistics, with the assumptions that the data utilized has been corrected and verified against 3D protein structures, and it makes no attempt to be robust against systematic referencing issues in the SHIFTX method. When analyzing experimental data, it was previously necessary to apply a recursive approach: define a raw reference value; perform the downstream analysis, refine the reference; and repeat the process. Considering these potential artifacts, the statistics that we employed cannot always be directly equated to the true chemical shift statistics of the amino acids present in assigned proteins. Also, RefDB only utilizes chemical shift datasets from proteins with well-defined structure, which means that the BaMORC algorithm is likewise tuned for such datasets.

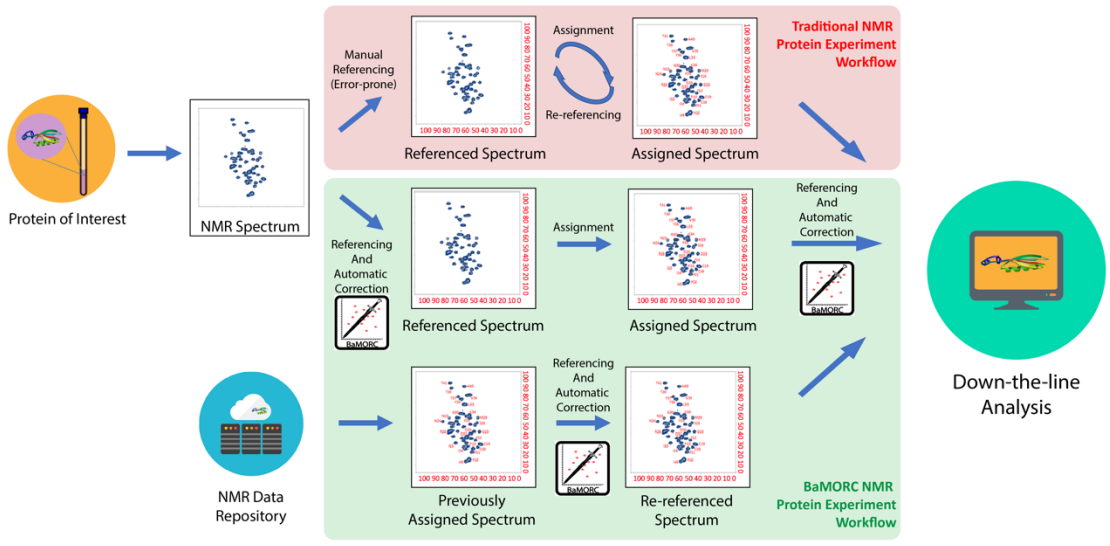

Figure 4.1 Overview of the traditional versus the BaMORC protein NMR reference correction workflows.

4.3 Design of core algorithmic components

At the center of the project, a reference correction value was calculated through an optimization that minimizes the difference between the estimated amino acid frequencies through statistical modeling and the actual frequencies based on the amino acid sequence.

4.3.1 Calculation of protein amino acid frequency with secondary structure

The actual protein amino acid frequencies could be calculated from the counting of each amino acid in three secondary structures and divided by the total number of residues in the sequence as shown here; $AA\ Freq_{aa,ss} = \frac{AA\ Count_{aa,ss}}{\sum AA\ Count_{aa,ss}}$, where aa stand for each of the 19 amino acids that do not include glycine.

For example, for the following protein sequence and its accompanying secondary structure, we can calculate the amino acid frequencies given each of the secondary structures. And this procedure is illustrated in the Figure 4.2 and Table 2.1. In the Figure 4.2, we are showing the top line is the protein sequence and the bottom line is the residue-

wise secondary structure. Using the formula mentioned above, we can calculate the amino acid frequency give each of the secondary structures as showing in Table 4.1.

```
M-Q-V-W-P-I-E-G-I-K-K-F-E-T-L-S-Y-L-P-P-L-T-V-E-D-L-L-K-Q-I
C-C-C-C-C-C-C-C-C-C-C-C-B-B-C-C-C-C-C-C-C-H-H-H-H-H-H-C-C
```

Figure 4.2 Example hypothetical protein sequence and its corresponding secondary structure

Table 4.1 Amio acid frequence give secondary structure.

AA_SS	Count	Frequency
D-H	1	0.03448276
E-C	2	0.06896552
E-H	1	0.03448276
F-C	1	0.03448276
I-C	3	0.10344828
K-C	2	0.06896552
K-H	1	0.03448276
L-B	1	0.03448276
L-C	2	0.06896552
L-H	2	0.06896552
M-C	1	0.03448276
P-C	3	0.10344828
Q-C	2	0.06896552
S-C	1	0.03448276
T-B	1	0.03448276
T-C	1	0.03448276
V-C	1	0.03448276
V-H	1	0.03448276
W-C	1	0.03448276
Y-C	1	0.03448276
Total	29	1

4.3.2 Predicting secondary structure using JPred

If the secondary structure information isn't given, which is the most common case in real-world protein NMR analysis, many secondary structure prediction methods are

available for this purpose. After comparing and testing many secondary structure prediction packages available, we identified JPred[97] as the general best one for our purposes based mainly on accuracy, but also general availability and the level that the method is maintained. Since 1988, JPred, a protein secondary structure prediction server has been operating and providing accurate prediction of residue-wise secondary structure from protein sequence. Behind the scenes, JPred utilizes the Jnet algorithm[98], which uses a neural network secondary structure prediction algorithm with different type of multiple sequence alignment profiles derived from the same sequence.

To fetch the secondary structure predictions, we have developed a JPred fetcher function for this very purpose, based on the provided instructions for the JPred web service. The JPred fetcher function submits a protein sequence to the server, which returns a unique job ID. Then using the job ID, the secondary structure predictions are downloaded when the JPred analysis is complete. Next using the same approach mentioned in 4.3.1, the amino acid frequencies can be calculated.

4.3.3 Estimation of protein amino acid frequencies using statistical modeling

Modeling the protein amino acid frequency is through the calculation of the density from a chi-squared distribution given the alpha and beta carbon chemical shifts. Assuming each pair of C_α and C_β chemical shifts follows a chi-squared distribution (X^2) with two degrees of freedom, we can calculate amino acid probabilities or density for each secondary structure using the following Bayesian formula: $P(AA_i|CS) = \frac{P(CS|AA)\times P(AA)}{\sum(P(CS|AA_i)\times P(AA_i))}$, where CS is chemical shifts, AA is amino acid, and $P(\cdot)$ is the probability.

Then we sum all the probabilities for each amino acid to calculate amino acid probability frequencies: $AAProbCount_i = \sum_i P(AA_i|CS_i)$, and $AAProbFreq_i = \frac{AAProbCount_i}{\sum AAProbCount_i}$.

4.4 Optimization to minimize differences between predicted and actual amino acid frequencies.

As with all the statistical learning or machine learning methods, the reference correction value is calculated through an optimization process, in this case by minimizing the difference between predicted and actual amino acid frequencies. For describing this optimization, we start with the L-1 or L-2 norms defined as: $\|x\|_1 = \sum_i |x_i|$ or $\|x\|_2 = \sqrt{\sum_i x_i^2}$.

The difference between the L-1 and L-2 norms can be understood geometrically. The L-2 norm is a form of least squares and easier to understand since it minimizes a Euclidean distance. The L-1 norm (referred to as the Manhattan or the Taxicab norm) represents the distance between two points by using the sum of the absolute difference of their Cartesian coordinates.

In our optimization, mean-absolute error (MAE) and mean-squared error (MSE) are based on the L-1 and L-2 norms respectively: $MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$ and $MSE = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$

In the final algorithm, we used MSE in the optimization, since an L-2 norm more often provides a single global minimum, while an L-1 norm and thus MAE more often provide multiple minima, which will complicate the optimization procedure.

4.5 BaMORC algorithm overview

Figure 4.3 provides a simplified overview of the overall BaMORC algorithm. With the protein sequence and secondary structure, predicted by JPred if not provided, we can calculate the actual amino acid and secondary structure composition, i.e. amino acid frequencies give secondary structure. Using the chi-squared density function with given statistics of each amino acid and secondary structure, we can estimate the amino acid and secondary structure composition from the alpha and beta carbon chemical shifts. Then by calculating the MSE between these two compositions, updating the reference value and repeating same procedure in iterative fashion, we can eventually find the best reference value that give us the smallest MSE. This best value is the final reference correction value reported.

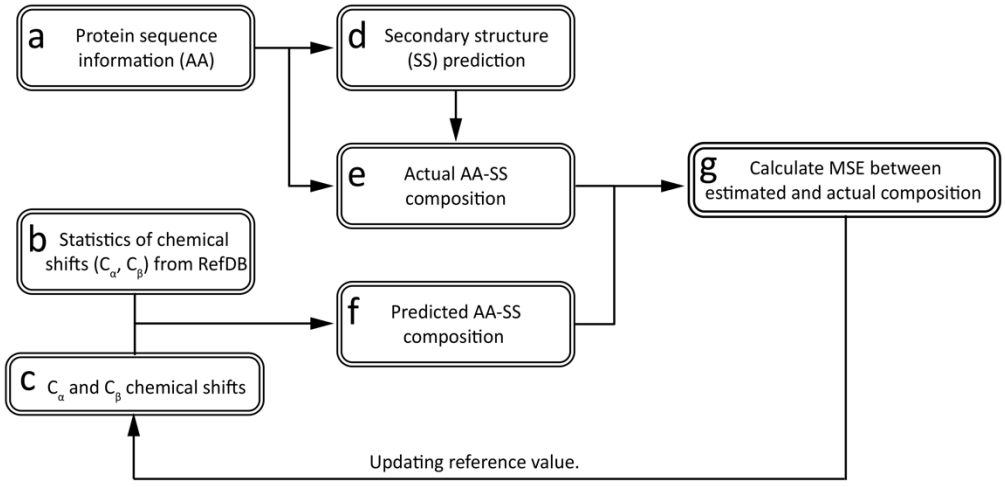

Figure 4.3 Overview of the project.

CHAPTER 5. BAMORC—TOOL FOR PROTEIN NMR REFERENCE CORRECTION

5.1 Introduction

Our algorithm named Bayesian Model Optimized Reference Correction (BaMORC) can detect and correct ^{13}C chemical shift referencing errors before the protein resonance assignment step of analysis and without three-dimensional structure. By combining the BaMORC methodology with a new intra-peaklist grouping algorithm, we created a combined method called Unassigned BaMORC that utilizes only unassigned experimental peak lists and the amino acid sequence [57,82]. Unassigned BaMORC kept all experimental three-dimensional HN(CO)CACB-type peak lists tested within +/- 0.4 ppm of the correct ^{13}C reference value. On a much larger unassigned chemical shift test set, the base method kept ^{13}C chemical shift referencing errors to within +/- 0.45 ppm at a 90% confidence interval. With chemical shift assignments, Assigned BaMORC can detect and correct ^{13}C chemical shift referencing errors to within +/- 0.22 at a 90% confidence interval. Therefore, Unassigned BaMORC can correct ^{13}C chemical shift referencing errors when it will have the most impact, right before protein resonance assignment and other downstream analyses are started. After assignment, chemical shift reference correction can be further refined with Assigned BaMORC. These new methods will allow non-NMR experts to detect and correct ^{13}C referencing error at critical early data analysis steps, lowering the bar of NMR expertise required for effective protein NMR analysis.

5.1.1 Calculating the overlap matrix and classifier weights

Sixteen of the 19 amino acid C_α-C_β bivariate distributions overlap almost completely, as shown in Figure 3.7. Due to the linearity of the statistical model, our methodology will favor those amino acid and secondary structure types with broad

distributions and lead to over-prediction of those types. To side-step this problem, we applied a Bayesian-inspired reverse logic approach on top of the traditional statistical model. Starting with the traditional model, we use the data X, i.e. the C_α and C_β chemical shift values, to calculate Y', the normalized amino acid and secondary structure probability sums, which represents an estimate of the amino acid and secondary structure composition. Y is the normalized amino acid and secondary structure frequencies. To calculate the reference value in a traditional manner, the difference between Y' and Y, calculated by the sum of the absolute or squared difference, is minimized by a grid-search of possible reference values. However, to deal with the overlapping properties of the amino acid distribution, we then multiply Y by the probability overlap matrix to calculate $\widehat{Y'}$, which is substituted into the difference calculation. Therefore, we end up minimizing the difference between Y' and $\widehat{Y'}$ instead, thereby turning a discrete classification into a "fuzzy" classification and capturing the overlap characteristics of the data. This algorithm is an adaptation of the adversarial approach[99]. With an image recognition example, the computer recognizes a generated image (I_{gen}) by comparing it with the actual image (I_{act}). The common approach is to minimize the difference between I_{gen} and I_{act}; however, the image generator function does a poor job due to limits in resolution. To help the computer out, we can use a "fuzzy" or downscale filter and apply it to the actual image, $I_{fuzzy} = f_{filter}(I_{act})$. Then the computer will have a better chance to recognize the actual image by comparing between I_{fuzzy} and I_{gen}, as shown in Figure 5.1.

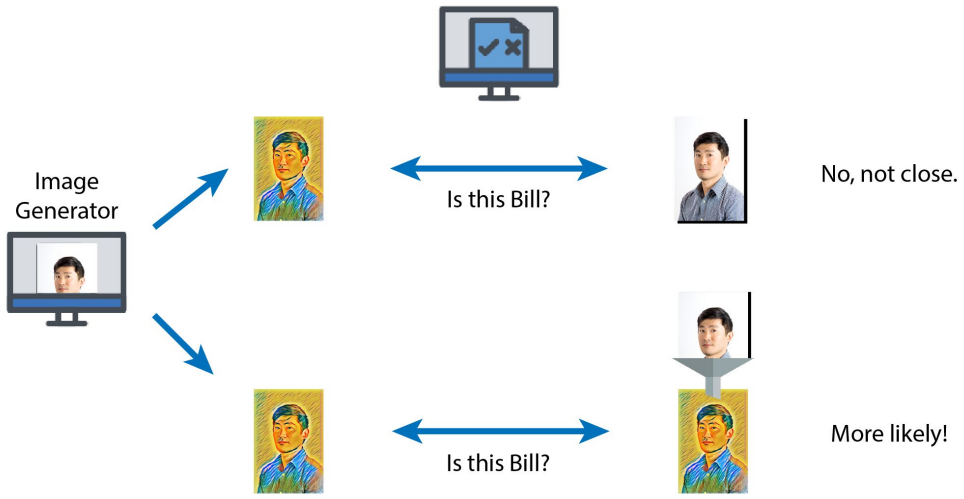

Figure 5.1 Overlapping matrix application rationale. Using a filter to bias the true image, will help computer to recognizing the correct answers.

Similarly, to calculate the $\widehat{Y'}$, we used the following equation: $\widehat{Y'} = Y \times \Omega_{overlap}$. Since we considered three secondary structure types here, the dimensions of both $\widehat{Y'}$ and Y were 1×57, and the $\Omega_{overlap}$ is a 57×57 matrix. When considering glycine, a 3×3 overlap matrix was employed. Finally, we concatenated the three glycine results into the 57-element vector to form a new $\widehat{Y'}$ and Y with 1×60 dimensions. The prediction overlap matrix calculation is based on probability calculations derived from each of the 60 statistical models. On the basis of amino acid types (excluding glycine) and secondary structure, we first grouped all of the chemical shifts into 57 bivariate groups/classes and 3 univariate groups/classes for glycine. Then, for every pair of C_α and C_β chemical shifts, we calculated the probabilities of the 57 classes. Likewise, we used every glycine C_α chemical shift to calculate the probabilities for the 3 glycine classes. For example, for every data point of an alanine-beta strand, we calculated the probabilities of all of the

classes. Then, we performed normalization across the columns and finally obtained a $57 \times 57_{57\times57}$ matrix.

In nature, amino acid chemical shift distributions are not ideal; i.e., the C_α/C_β bivariate statistical models approximate the real distributions. Hence, we used the real distributions to calculate the prediction overlap between the bivariate statistical models and represented this overlap as prior information in the form of a prediction overlap matrix. Moreover, we employed the diagonal elements of this matrix (Figure 5.2) as weights (ω_i's), in the calculation of residuals. Top figure of Figure 5.2 shows Probability overlapping matrices for amino acids excluding glycine and bottom, the overlapping matrix for glycine. The color represents the value in the matrix: a higher value corresponds to a darker red color, and a lower value to a light yellow. Higher diagonal probabilities indicate better predictive power of the given model. This maximizes the use of classifiers with the least overlap and, thus, the best prediction performance.

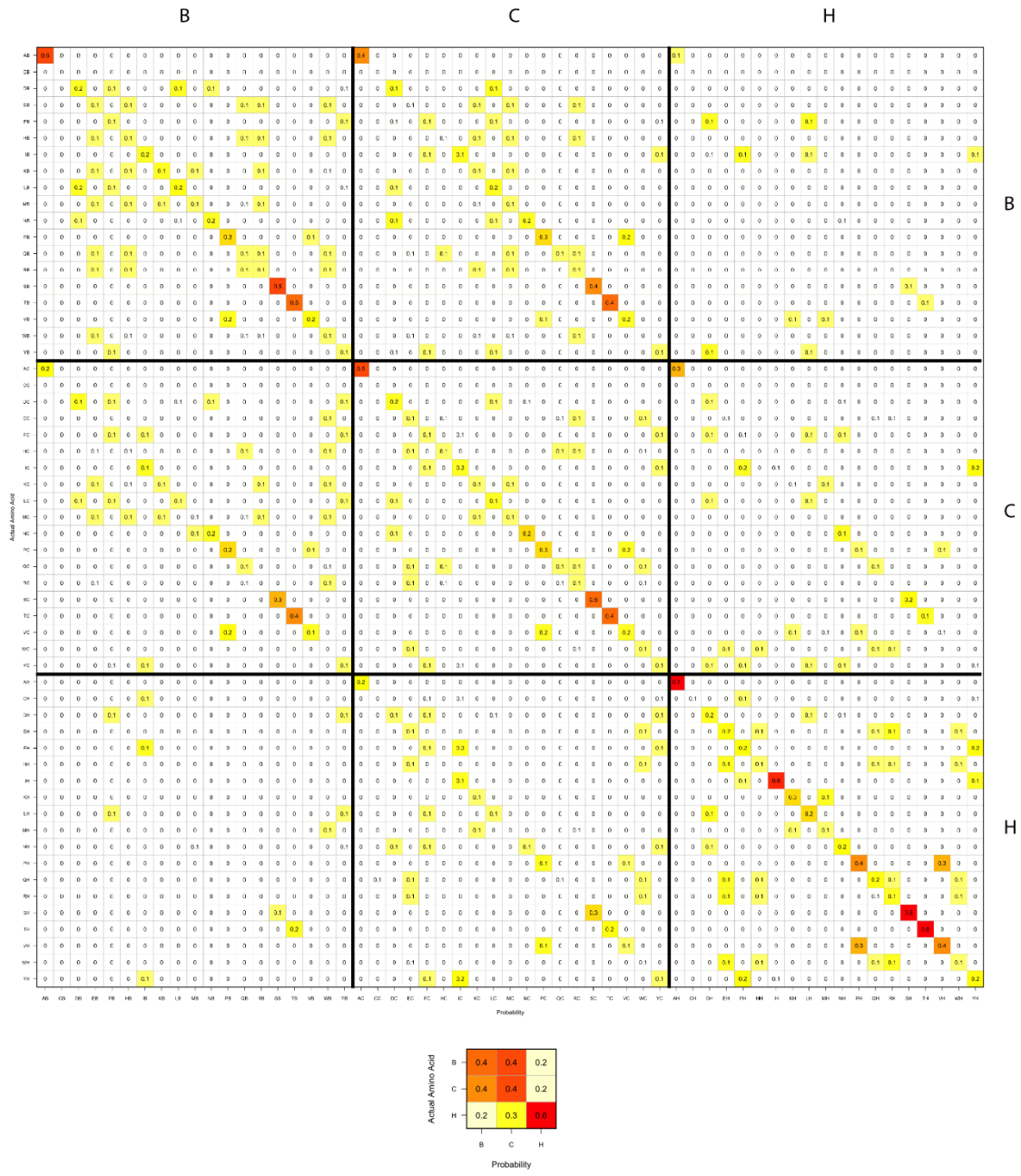

Figure 5.2 Bayesian prediction overlap prior matrix derived from the bivariate statistical models and chemical shifts from the RefDB.

The overall optimization approach can be simplified into the following residual equation which is minimized as showing here: $\min\left(\sum \omega_i \left|Y_i' - \widehat{Y_i'}\right|\right) = \min\left(\omega \cdot \left|Y' - Y \cdot \Omega_{overlap}\right|\right) = \min\left(\left|\omega \cdot Y' - \omega \cdot Y \cdot \Omega_{overlap}\right|\right)$.

To calculate the $\widehat{Y'}$, we multiplied Y, the ground truth, with the overlap matrix, $\Omega_{overlap}$. This $\widehat{Y'}$ captures the overlap characteristics of the statistical models with respect to the data. Then, to best utilize the statistical models with the best predictive power, we further multiplied Y' and $\widehat{Y'}$ by the weights, ω. By utilizing a grid-searching method, we identify an optimal value that minimizes the absolute difference between the outcomes from both the estimated and actual amino acid and secondary structure compositions.

5.2 BaMORC methodology

The bottom right flowchart in Figure 5.3 provides an overview of the BaMORC method. In describing this method, let $V_{AA,SS}$ denote the_ chemical shifts space. $V_{AA,SS} = (X_{c_a,1}, X_{c_b,1}),\ \dots\ , (X_{c_a,2}, X_{c_b,2}) V_{AA,SS} = (X_{C_\alpha,1}, X_{C_\beta,1}), \dots (X_{C_\alpha,n}, X_{C_\beta,n})$, where $AA \in (19\ amino\ acid\ typings) AA \in (19\ Amino\ Acid\ Typing)$ and $SS \in (3\ secondary\ structure\ types) SS \in (3\ secondary\ structure)$.

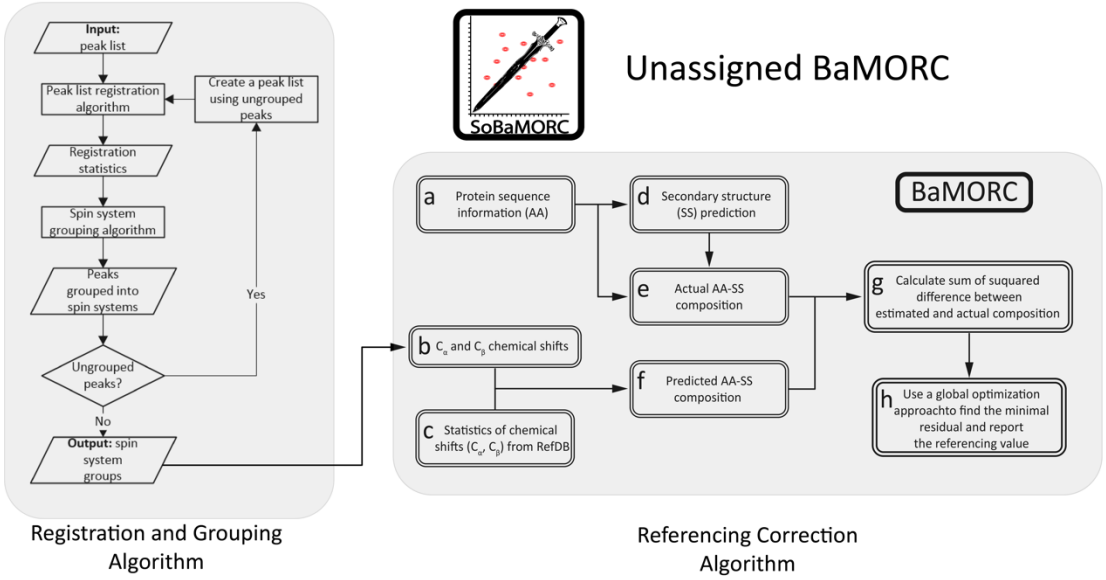

Figure 5.3 Flow diagram of the Assigned and Unassigned BaMORC method.

We exclude glycine here for simplicity, since it does not have a beta carbon. The reference correction method assumes that for each $V_{AA,SS}$, it follows a unique bivariate normal distribution. For example, $V_{Alanine,Helix} \sim MVN(\mu_{c_a,A,H}, \mu_{c_b,A,H}, \Sigma_{A,H})$

$V_{Alanine,Helix} \sim MVN(\mu_{c_a,A,H}, \mu_{c_b,A,H}, \Sigma_{A,H},)$, whereby a covariance ($\Sigma$) exists between and the α and β ^{13}C chemical shifts. To calculate the probability, we first need to transform each pair of the chemical shifts to a chi-square value using equation $\chi^* = \left[v - \left(\hat{\mu}_{c_a,AA,SS,} \hat{\mu}_{c_\beta,AA,SS,}\right)\right] \times \Sigma^{-1}_{AA,SS} \times \left[v - \left(\hat{\mu}_{c_a,AA,SS,} \hat{\mu}_{c_\beta,AA,SS,}\right)\right]^T$

$\chi^* = [v\text{-}(\hat{\mu}_{c_a,AA,SS}, \hat{\mu}_{c_b,AA,SS}))] \times \hat{\Sigma}^{-1}_{AA,SS} \times [v\text{-}(\hat{\mu}_{c_a,AA,SS}, \hat{\mu}_{c_b,AA,SS}))]^T$, and χ^* follows a chi-square distribution with 2 degrees of freedom χ^2_2 (for glycine, χ^2_1). But in the final version of our method, we removed glycine models based on robustness testing. Then, we can calculate the probability of each of the amino acid type and secondary structures for any

pair of α_{α} and β ^{13}C chemical shifts. For a given NMR dataset with n pairs of chemical shifts, the BaMORC will calculate 57 possibilities for each pair of chemical shifts and 3 possibilities for single chemical shifts, among which the maximized probability represents the corresponding amino acid type and secondary structure. The BaMORC method computes every probability across the dataset, sums up them based on amino acid type and secondary structures, and then normalizes the sums so that the sum of the sums is equal to 1. These 57 sums represent the estimated composition frequency. The difference between the estimated composition and the actual composition, which is calculated from the sequence, is minimized via a grid search. The assumption is that the dataset with the correct reference should report the lowest difference, as the two compositions should match closely.

The search range is typically limited to -5 to 5 ppm centered around the current reference value of 0. The algorithm first evenly samples 50 candidate reference correction values in the range from -5 to +5. Each of the candidate values is applied in the whole dataset, and the difference between the estimated and actual amino acid composition frequency is calculated. The one value that minimizes the difference is the raw correction value, M_1, and then around this value the algorithm will evenly sample another 50 candidates around this value, from the range between $M_1 - 1$ and $M_1 + 1$. The algorithm subsequently performs the same calculation to identify the value that minimizes the difference and reports it as the final correction value, M_2. To further reduce the computational time, we also utilized global optimization algorithm[100] to estimate the referencing correction value, and we will further describe it in 5.3.2.

5.2.1 Assigned BaMORC method

The assigned BaMORC approach uses the assigned amino acid type information along with secondary structure prediction from JPred to greatly reduce the number of amino acid typing probabilities that are calculated, i.e. from 60 probability calculations for each C_α/C_β pair in BaMORC to only 1 probability calculation (step f in Figure 5.3).

To further reduce the computation time and allow a better user experience, we exchanged the grid-search approach with a function from the Global Optimization by Differential Evolution (DEoptim) library[101] as shown in Figure 5.4. This global optimization function was implemented using the differential evolution algorithm (DE) [102]. Three max iteration parameters were used for DEoptim function: 10, 20, 50. The violin plots here show the distribution of the results. The mark on the top of each plot is the 95% quantile and the one on the bottom is the 5% quantile. The boxplots show the 75%, 50% and 25% quantiles respectively. The results from these three settings are very similar. With the higher iteration value, the results get better trivially. Round-up mean values are all 0.08 ppm, which is same the grid search algorithm. All of the DEoptim results have a 0.75 ppm range at the 90% confidence interval, which share the same trend of the mean values, and they are different from grid search by 0.05 ppm range at the 90% confidence interval.

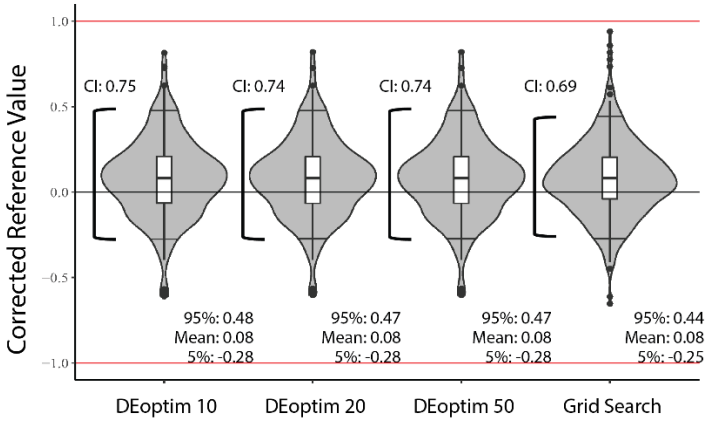

Figure 5.4 Comparison of BaMORC performance using grid search optimization vs global optimization by differential evolution.

We tested three max iteration numbers for the global optimization DEoptim function: 10, 20, 50. The results from these three settings are very similar, with the hither iteration value, the results get better trivially but computational time increase exponentially, which is from >2 minutes to <15 minutes per dataset. Also, the resulting optimization problem has a smooth enough error surface to use better optimization methods than a grid search. Therefore, we included the global optimization by differential evolution (DEoptim) [100,101]. Both improvements together decrease the running time of the method to less than 1 minute. The comparison of BaMORC performances using grid search optimization vs global optimization was shown in Figure 5.5. In essence, the global optimization computational timing is shorter with a better performance for NMR data with assignment results.

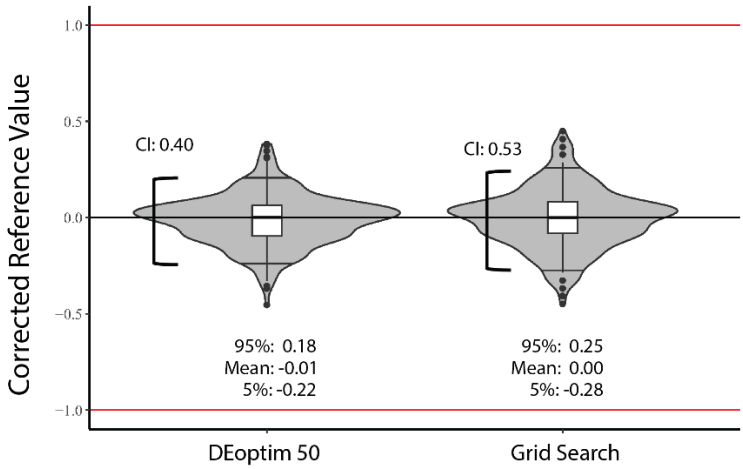

Figure 5.5 Comparison of Assigned BaMORC performance using grid search optimization vs global optimization by differential evolution.

The results from DEoptim function, with the max iteration value equal to 50, performs much better than our original grid search implementation. The violin plots here show the distribution of the results. The mark on the top of each plot is the 95% quantile and the one on the bottom is the 5% quantile. The boxplots show the 75%, 50% and 25% quantiles respectively. The mean correction values are -0.01 and 0.00 respectively. The DEoptim results have a 0.40 ppm range at the 90% confidence interval, which share the significantly different from grid search by 0.13 ppm range at the 90% confidence interval.

5.2.2 Unassigned BaMORC Method

Conceptually, the algorithm consists of two parts. A full schematic representation of the analysis workflow is provided in Figure 5.3. The first part of the Unassigned BaMORC method groups the peaks in the 3D HN(CO)CACB peak list into spin systems using ^{1}H and ^{15}N common resonances[28]. Ideally, the HN(CO)CACB peak list will contain two peaks for every amino acid except for glycine, which lacks a beta carbon, so the number of spin system groups in the HN(CO)CACB peak list will be equal to the number

of amino acids minus the number of glycine residues. The second part of the Unassigned BaMORC method uses the $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ carbons chemical shifts for every spin system group returned by the grouping algorithm and employs the BaMORC method to calculate and return the carbon reference correction value.

Grouping methodology (spin system grouping algorithm). The spin system grouping algorithm, as illustrated in Figure 5.3, can group peaks into spin systems in peak lists that have multiple peaks per spin system. In this use-case, the HN(CO)CACB NMR peak list contains 2 peaks for each spin system group except for the glycine residues. The grouping of peaks into spin systems is complicated by the presence of multiple sources of variance in dimension-specific peak positions; i.e., different dimension-specific match tolerance values are necessary to reliably group peaks into spin systems without overlap. Our grouping algorithm consists of two parts: the registration step and the actual grouping step[28]. The registration step derives the necessary match tolerance values from the single-peak lists necessary to group peaks into spin systems. The grouping algorithm is based on the widely-used density-based clustering algorithm DBSCAN[103] and employs derived dimensions-specific match tolerances values to group peaks into spin systems. It uses a chi-squared distance cutoff and variance-normalized distance (chi-square value) to decide whether the peaks can be grouped into spin systems. To address the problem of multiple sources of variance, the algorithm is developed in an iterative fashion, which allows it to readjust match tolerance values in the case where peaks are left ungrouped by repeating the registration step again and grouping as many peaks into spin systems as possible. Figure 5.3 is the flow diagram of the iterative grouping algorithm. First, the grouping algorithm reads in a single peak list in and runs the registration in order to identify the

initial match tolerances for each comparable dimension (for ^{1}H and ^{15}N in the case of HN(CO)CACB). Next, it groups peaks into spi system clusters using the derived match tolerance values. Then, the algorithm checks whether any ungrouped peaks remain and, if so, creates a new peak list and attempts to register it again itself again to determine new, larger match tolerances that can be used to group peaks into spin systems.

Reference correction methodology (BaMORC). The reference correction methodology is essentially BaMORC. The input of the algorithm is the output from the grouping methodology, which are pairs of ^{13}C chemical shifts derived from pairs of grouped HN(CO)CACB peaks. Using these pairs of ^{13}C chemical shifts and the same BaMORC analysis pipeline reports an optimized correction value as a reference. Eventually, Unassigned BaMORC applies this correction value to all of the C_α and C_β chemical shifts and prints out a text file, that including all of the corrected peak lists in the final output.

5.3 Results

5.3.1 Initial evaluation of different covariance statistical models for unassigned NMR reference correction

We created an unordered pair of C_α and C_β chemical shifts for a given residue, which we will refer to as a carbon spin system in this context. Unordered pairs were used to test the situation where the amino acid assignment of chemical shifts is not known. Five types of covariance matrices, represented by Matrices A-E, were tested under a generalized chi-squared method to calculate the chemical shift probabilities for each carbon spin system within the BaMORC methodology (see 5.2). The calculation of variances (sd^2) and

covariances (Cov) are described in Equations 6-8 in the Methods section. Matrix E utilizes the full set of amino acid- and secondary structure-specific covariances. As mentioned previously, we discovered that the majority of the RefDB datasets are from multiple NMR experiments and are not appropriate for extracting covariance statistics. As described in the Methods, we used the RMSD values of each dataset as a criterion to further filter out datasets that are likely not derived from a single NMR experiment and to develop the Matrix E-revised method. In Figure 5.6, we show the results across different methods using all the RefDB data. Across all of the RefDB data, E-Revised covariance matrix calculated from filtered data performed better. The violin plots here show the distribution of the results. The mark on the top of each plot is the 95% quantile and the one on the bottom is the 5% quantile. The boxplots show the 75%, 50% and 25% quantiles respectively. With both the E-Revised covariance matrix and the Bayesian prediction overlap matrix prior, the algorithm performs the best. Covariance matrices A and C perform similarly, with 90% interquartile ranges (IQRs) of 2.37 and 1.80. Covariance matrices B, D and E show worst results, since their means deviate greatly from the true reference value. The E-Revised matrix performs better, with a 90% IQR of 1.35 and a mean of -0.20, which is very close to the true reference. After applying the Bayesian prior prediction overlap matrix, the performance of BaMORC shows a dramatic improvement, with a 90% IQR of 0.73 and mean of -0.08, which far outperforms the state-of-the-art algorithms. When applying the same algorithms on the data with at least 90% completion, the performance of BaMORC remains stable with small improvement, with a 90% IQR of 0.69 and same mean of -0.08.

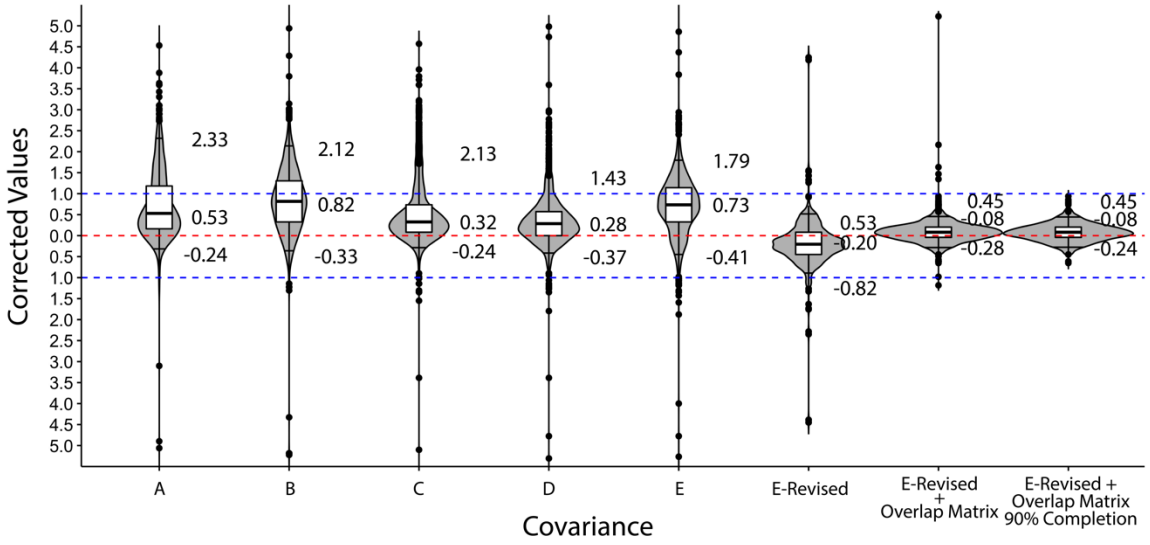

Figure 5.6 Results across different methods using all RefDB data.

The violin plots in Figure 5.6, similar to box plots, but with a visual representation of the full distribution (i.e. a sideways, mirrored histogram), illustrate that the initial Matrix E, which incorporates three separate secondary structure covariances, does not perform well as compared to Matrix D (including the averaged covariance of three secondary structure) and to Matrix B (including no covariance information). The performance is measured on the y-axis (Corrected Reference Value) as a comparison to 568 RefDB datasets treated as a gold standard. This poor performance is due to the use of inaccurate covariances arising from the inclusion of entries that lack the correct correlation between C_α and C_β chemical shifts, since these shifts may come from separate spectral sources. Matrix E-Revised showed the best performance among the pure statistical models, exhibiting the closest ^{13}C reference correction of 0.00 ppm for BMR6032 entry, as shown in Figure 5.7 and Table 5.1. The performance of Matrix E-Revised as illustrated in Figure 5.7 demonstrates the significant improvement in predictions that even small changes in covariances can provide. In addition, for the BMR6032 entry in Figure 5.7, both the shape

of the penalty function to be minimized and the overall minimum value are affected by the type of covariance matrix: all panels show step-wise plots of the second 50-step grid search, with the corresponding covariance matrix presented below. The covariance matrices A, C and E all show a major deviance from the true reference value. Matrices B and D perform equally well but with small deviance. The E-Revised matrix performs the best, with its output exactly matching the true reference value.

Table 5.1 Performance of different covariance matrices on the BMR6032 dataset

Covariance	Estimated Reference Value (ppm)
A	0.327
B	-0.0408
C	0.245
D	-0.0408
E	0.327
E-Revised	0.000

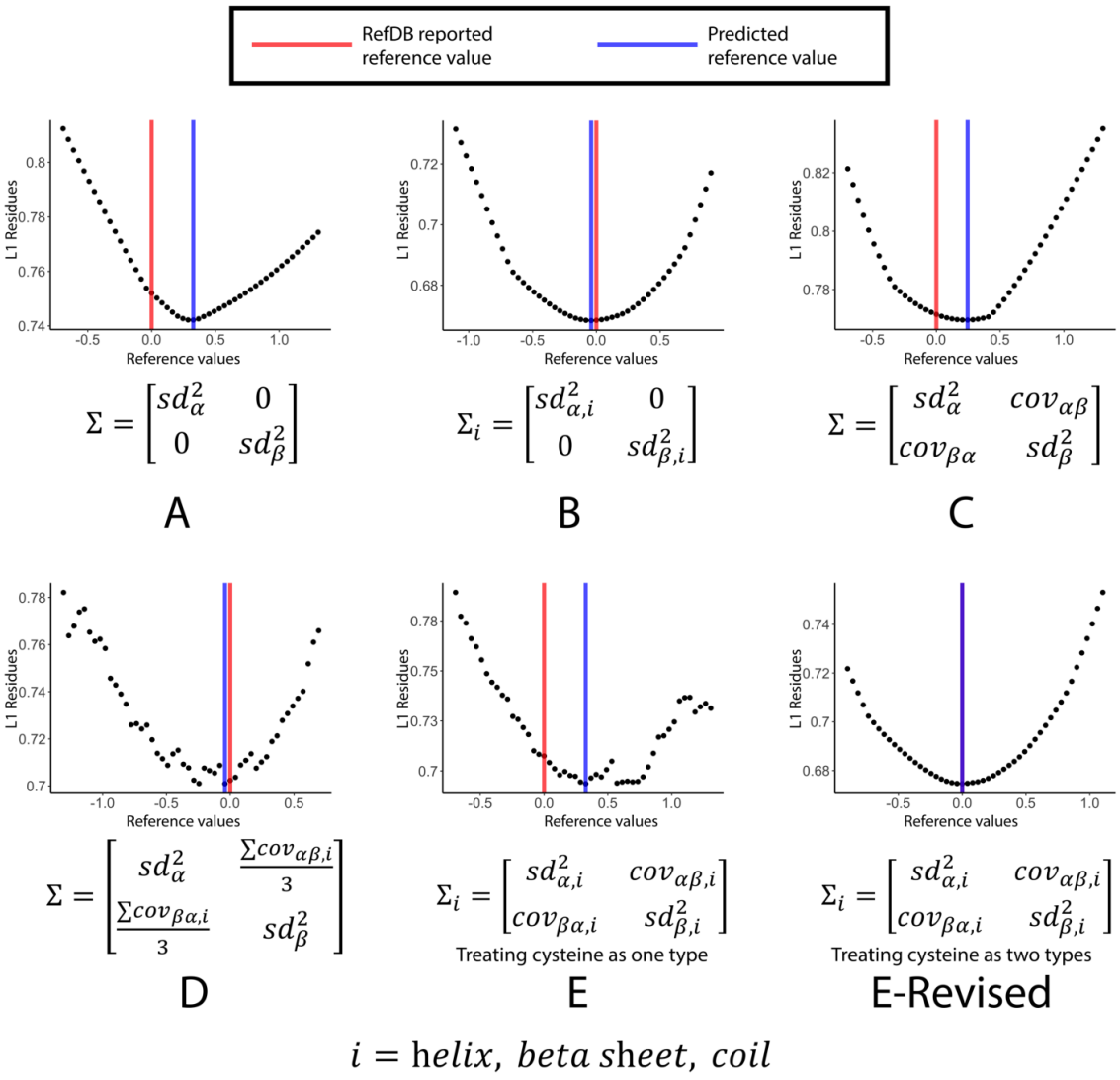

Figure 5.7 Performance of different covariance matrices on the BMR6032 data.

5.3.2 Correcting for overlap in amino acid type predictions between statistical models

As Figure 3.7 illustrates the substantial overlap of bivariate distributions for a majority of the amino acids. Most statistical learning (SL) algorithms will be biased in favor of certain amino acid types with broad distributions, leading to inaccurate prediction of amino acid and secondary structure types. The standard SL approach estimates an amino acid content frequency (Y') that is close to the observed amino acid content frequencies

(Y) via minimizing the difference between Y' and Y through specific optimization or search procedures. However, due to the linear relationship limitation, the estimated result Y' can never eliminate the effects of overlap observed in the amino acid- and secondary structure-specific bivariate distributions in the NMR data. Therefore, we applied a Bayesian-inspired reverse logic to estimate the overlap effects of the C_α/C_β bivariate statistical models on the observed amino acid content frequencies Y in order to produce $\widehat{Y'}$. This is accomplished by generating a prediction overlap matrix from the estimated frequency of overlap across C_α/C_β bivariate statistical models using observed C_α/C_β chemical shifts in the RefDB associated with specific amino acid and secondary structure types. The observed amino acid content frequencies Y is multiplied by the resulting prediction overlap matrix to produce $\widehat{Y'}$, which mimics the effects of overlap. As an analogy, paper turns yellow from the effects of aging. This aging effect can be mimicked by staining a new piece of paper with tea or coffee and then heating the paper to turn it yellow and make it appear to be old. Likewise, the prediction overlap matrix is mimicking the effects of overlap caused by the statistical modeling. In other words, the prediction overlapping matrix acts like a Bayesian prior in estimating the effect of overlap on the observed amino acid content frequencies Y. This Bayesian-inspired approach is illustrated in Figure 5.8 and detailed in the Methods session. Figure 5.7 shows the prediction overlap matrices for all 20 amino acids. We also employed the diagonal elements of the prediction overlap matrix as weights in the comparison and minimization of differences between Y' and $\widehat{Y'}$. Thus, the comparison of Y' and $\widehat{Y'}$ utilizes the most discriminating predictors based on prediction accuracy and on the observed prevalence of C_α and C_β chemical shifts in real

datasets. The calculation of the prediction overlap matrix and predictor weights is described in the Methods.

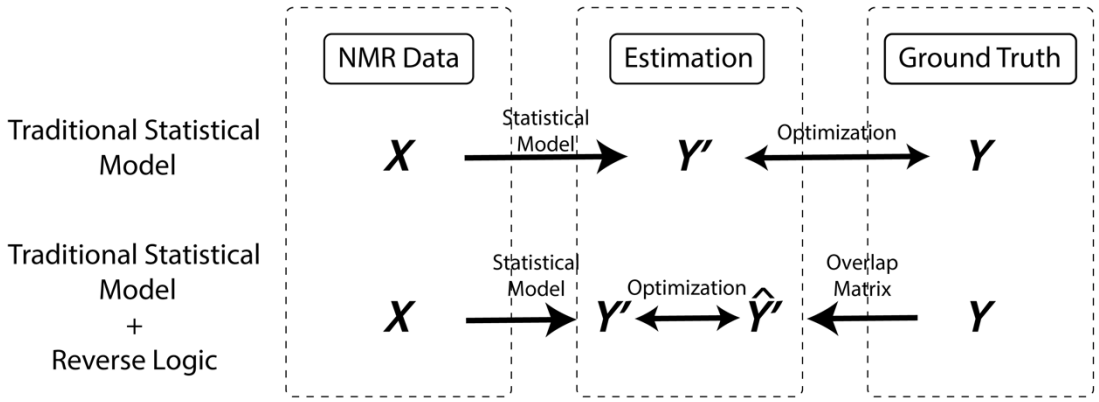

Figure 5.8 The BaMORC approach with a Bayesian prediction overlap prior matrix.

The BaMORC method combines the E-Revised covariance method used in the chi-squared-based C_α/C_β bivariate statistical models with the prediction overlap matrix, while ignoring glycine residues. The BaMORC method improves the comparison of the predicted and observed amino acid and secondary structure frequencies more than 2.5-fold by modifying Y with the prediction overlap matrix to create $\widehat{Y'}$, which reflects the overlap introduced by Matrix E-Revised. All of the other statistical models were also tested but performed significantly worse than the BaMORC method, as illustrated by the violin plots in Figure 5.6 and Table 5.2.

Table 5.2 Quantiles and IRQs results from a series of statistical models tested against all of the data from the RefDB

Covariance	5% (ppm)	25% (ppm)	50% (ppm)	75% (ppm)	95% (ppm)	90% IQR	50% IQR
A	-0.24	0.16	0.53	1.18	2.33	2.57	1.02
B	-0.33	0.33	0.82	1.3	2.12	2.94	0.97
C	-0.24	0.08	0.32	0.73	2.13	2.37	0.65
D	-0.37	0	0.28	0.57	1.43	1.8	0.57
E	-0.41	0.33	0.73	1.14	1.79	2.2	0.81
E-Revised	-0.82	-0.41	-0.2	0.04	0.53	1.35	0.45
E-Revised + Overlap Matrix	-0.28	-0.04	0.08	0.2	0.45	0.73	0.24
E-Revised + Overlap Matrix (90% Completion)	-0.24	-0.04	0.08	0.2	0.45	0.69	0.24

In Figure 5.6, we compared the results of reference correction from the set of statistical models based on each covariance matrix (A-E, E-Revised) and the E-Revised covariance matrix with the prediction overlap matrix as applied to all the unassigned RefDB datasets. In this comparison, the E-Revised covariance matrix combined with the prediction overlap matrix acting as a Bayesian prior demonstrated overwhelming performance. The 90% confidence interval was +/- 0.45 ppm with an absolute length of 0.73 ppm. When we applied the same approach to the data with at least 90% completion, the BaMORC reference results remain stable with small improvement.

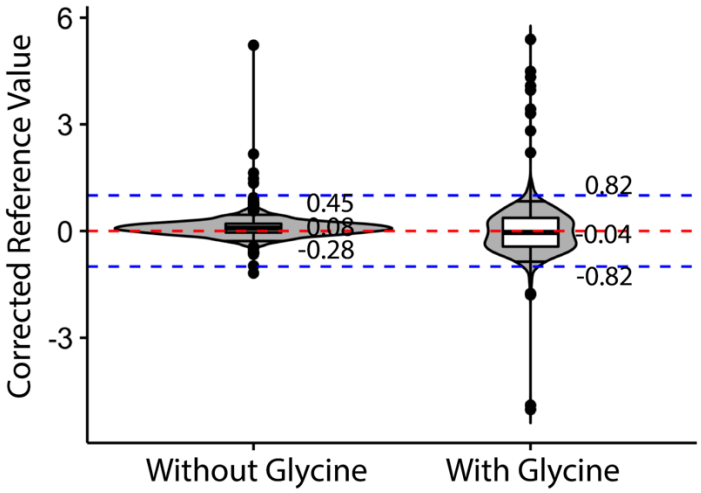

Figure 5.9 Performance of BaMORC methodology with and without glycine

We also tried to add glycine-specific predictors in the BaMORC method. However, the inclusion glycine statistical models had mediocre performance in comparison to using only the 57 non-glycine predictors. This is illustrated in Figure 5.9, which shows a bimodal distribution of reference correction values with a 90% confidence interval of +/- 0.82 ppm and absolute length of 1.64 ppm. Inclusion of glycine-specific statistical models had a worse performance than leaving these statistical models out of the full BaMORC method. The violin plots here show the distribution of the results. The mark on the top of each plot is the 95% quantile and the one on the bottom is the 5% quantile. The boxplots show the 75%, 50% and 25% quantiles respectively. The cause of the poor performance appears rooted in the complete overlap of C_α chemical shift distributions for beta sheet and coil secondary structure types for glycine residues. This is illustrated by the universally-high prediction-overlap values for glycine predictors as shown in Figure 5.2. The high values would significantly inflate the product of the matrix multiplication, which will greatly influence the residuals over the range of overlapping C_α chemical shift distributions. Thus, in the final implementation of the BaMORC methodology we ignored glycine residues.

5.3.3 Testing the robustness of the refined NMR shift reference correction method

Protein NMR datasets are typically incomplete from the perspective of what resonances are expected based on the protein sequence. This incompleteness is due to a host of experimental issues that prevent the detection of all protein resonances. In the RefDB itself, only 568 out of the 1557 entries include 90% or more of the expected C_α and C_β chemical shifts. Therefore, missing chemical shift data is a real issue that must be addressed. Accordingly, we tested the performance of the BaMORC method using unassigned datasets generated from the RefDB with varying amounts of missing ^{13}C spin systems. First, we constructed datasets with 100% completion by removing amino acid sequences for missing C_α and C_β chemical shift values for 568 entries with 95% or greater starting completion. Then, we incrementally removed 5% of the ^{13}C spin systems and tested the performance. Figure 5.10 and Supplemental Table 5.3 show the performance when 100% to 50% of ^{13}C spin system data are present. The overall performance of BaMORC does not appreciably deteriorate until approximately 70% of the ^{13}C spin systems were missing. Even, with 50% of the spin systems missing, the absolute length of the 90% confidence interval is less than 1 ppm, with the reference corrections within +/- 0.6 ppm. Therefore, BaMORC is very robust to missing ^{13}C chemical shift data.

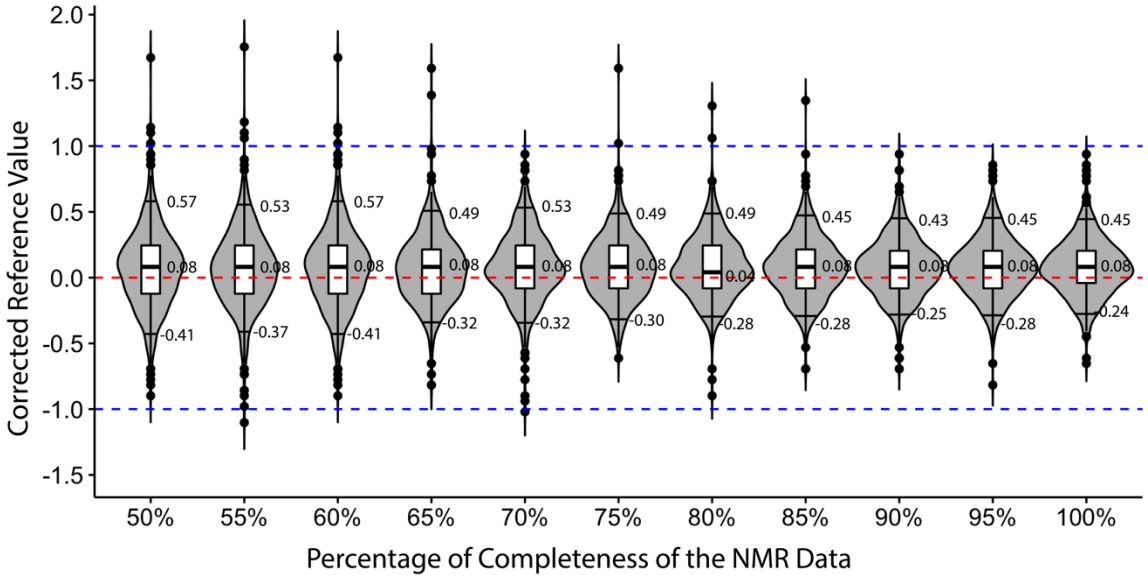

Figure 5.10 Testing the robustness of BaMORC against varying amounts of missing chemical shifts.

Table 5.3 Quantiles and IQRs for the robustness testing of the BaMORC method.

Completeness	5% (ppm)	25% (ppm)	50% (ppm)	75% (ppm)	95% (ppm)	90% IQR	50% IQR
100%	-0.24	-0.04	0.08	0.2	0.45	0.69	0.24
95%	-0.29	-0.08	0.08	0.2	0.45	0.69	0.24
90%	-0.24	-0.08	0.08	0.2	0.43	0.67	0.24
85%	-0.29	-0.08	0.08	0.21	0.45	0.69	0.26
80%	-0.29	-0.08	0.04	0.24	0.49	0.73	0.29
75%	-0.3	-0.08	0.08	0.24	0.49	0.73	0.29
70%	-0.33	-0.08	0.08	0.24	0.53	0.78	0.29
65%	-0.33	-0.12	0.08	0.21	0.49	0.73	0.26
60%	-0.41	-0.12	0.08	0.24	0.57	0.82	0.29
55%	-0.37	-0.12	0.08	0.24	0.53	0.78	0.29
50%	-0.41	-0.12	0.08	0.24	0.57	0.82	0.29

5.3.4 Testing BaMORC with predicted secondary structure

To test the performance of our method in a real-life situation, we removed all of the secondary structure information from the RefDB data and used the sequence-based secondary structure predictions generated from JPred4[97]. JPred4 is one of the best

algorithms for predicting secondary structure from sequence information alone, as showing in Figure 5.11 Performance (Matching Fraction) for JPred Algorithm on all RefDB datasets. We have tried other algorithm also, but JPred Algorithm gives us the best performance: 1258 out of 1557 datasets have a correct prediction percentage of over 70%. Across the RefDB, this breaks down to 46718 correct helix predictions out of 56015, 34063 correct coil predictions out of 73048, and 34063 correct beta strand predictions out of 50930. The new modified version of BaMORC performs as well with the JPred4 prediction as with the "true" secondary structure information from the RefDB, as summarized in Figure 5.12 and Table 5.4. This result may not be as surprising, since both the SHIFTX and JPred4 methods were developed from structure-based analyses.

Table 5.4 Quantiles and IQRs from the results of the BaMORC method performed using secondary structure information from RefDB and JPred.

Secondary Structure	5% (ppm)	25% (ppm)	50% (ppm)	75% (ppm)	90% (ppm)	90% IQR (ppm)	50% IQR (ppm)
RefDB	-0.24	-0.4	0.08	0.20	0.45	0.69	0.24
JPred	-0.24	-0.04	0.08	0.20	0.45	0.69	0.24

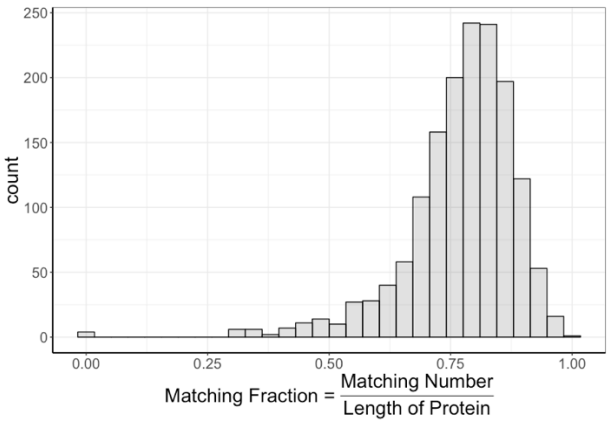

Figure 5.11 Performance (Matching Fraction) for JPred Algorithm on all RefDB datasets.

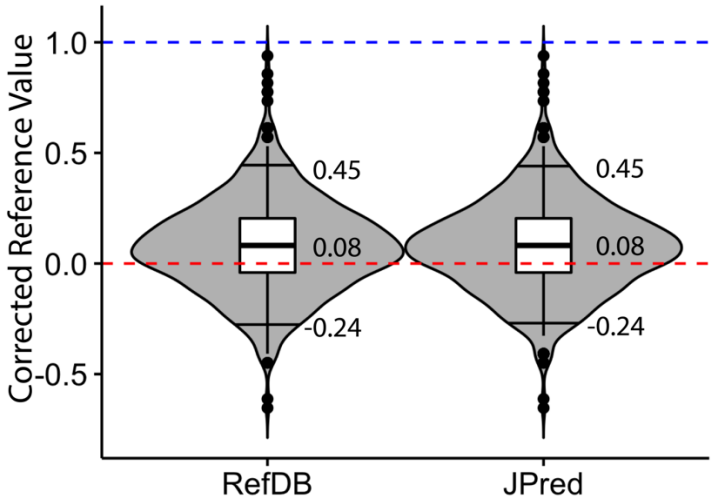

Figure 5.12 Comparison of the results obtained utilizing secondary structure information from RefDB and JPred4.

5.3.5 Testing assigned BaMORC versus LACS

While the BaMORC algorithm does not utilize assignment nor structure, we augmented and simplified the base algorithm to utilize assignment information in order to improve reference correction. This alternative implementation called Assigned BaMORC solves the same reference correction problem that the LACS method addresses. Assigned BaMORC takes an assigned NMR-STAR formatted file and returns a single reference offset/correction value for both alpha and beta carbons. We applied Assigned BaMORC and LACS to 1330 datasets from the RefDB with at least 90% assignment completion. On these datasets, assigned BaMORC outperformed LACS as shown in

Figure 5.13. Using known assignment, the Assigned BaMORC with DEoptim algorithm achieved much better results than the LACS algorithm. The violin plots here show the distribution of the results. The mark on the top of each plot is the 95% quantile and the one on the bottom is the 5% quantile. The boxplots show the 75%, 50% and 25% quantiles respectively. The results of Assigned BaMORC (left), it achieved a 0.40 ppm

range in confidence interval for data with 90% completion, while LACS achieve slight worse results 0.59 ppm range.

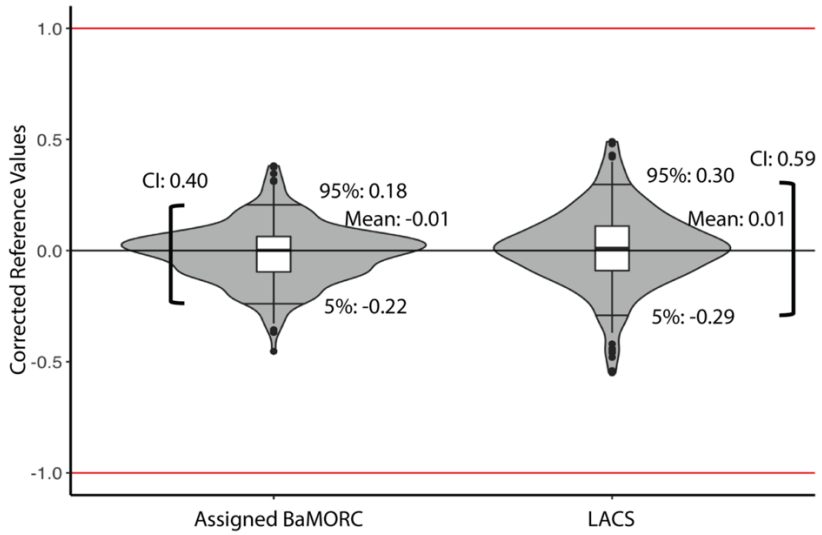

Figure 5.13 Comparison of Assigned BaMORC versus LACS performance on RefDB.

5.3.6 Testing unassigned BaMORC with experimental peak lists

In the case of real-world use, the data obtained from an NMR instrument are not labeled by resonance or grouped into spin systems. To further contribute to the protein NMR field, we applied a new intra-peak-list grouping algorithm developed in our laboratory [28] on top of the BaMORC method and developed a combined method, which we refer to as Unassigned BaMORC. This method can use unassigned three-dimensional HN(CO)CACB-type peak lists to correct the ^{13}C chemical shift referencing. This new tool greatly facilitates the automatic analysis and correction of NMR data before downstream analyses. Unassigned BaMORC generates a correction value, a file of re-referenced chemical shifts, and a residual plot showing the optimization of the predicted amino acid frequencies and where the best reference correction value occurs within the optimization. Table 5.5 shows the performance of Unassigned BaMORC on ten real peak lists derived

from solution NMR HN(CO)CACB spectra with secondary structure prediction provided by JPred. These peak lists were manually peak-picked. All ten experimental peak lists have Unassigned BaMORC-predicted reference correction values within +/- 0.40 ppm of the RefDB registration offset value, which is better performance than BaMORC's application across unassigned datasets derived from the RefDB. Two experimental peak lists from BPTI and Z domain of staphylococcal protein A have deviations greater than 2 ppm from the correct carbon chemical shift referencing. Also, none of these experimental peak lists are complete, with several peak lists having over 15% fewer spin systems than expected based on the protein sequence.

Table 5.5 Unassigned BaMORC performance with real-world examples.

Protein	Sequence length	Number of Spin Systems	BMRB ID PDB ID	RefDB Registration Offset Value	Unassigned BaMORC Reference Correction Value	Absolute Difference between Unassigned BaMORC and RefDB
Bovine pancreatic trypsin inhibitor (BPTI)[16]	58	47	5359 / 5PTI	-8.15	-8.55	0.40
Cold shock protein (CspA) [104]	70	57	4296 / 3MEF	-0.06	0.00	0.06
Protein yggU from E.coli (Target ER14) [105]	108	93	5596 / 1N91	-0.11	-0.20	0.09
Fibroblast growth factor (FGF)[106]	154	128	4091 / 1BLD	0.21	0.45	0.24
30S ribosomal protein S28E from Pyrococcus horikoshii (Target JR19)[107]	82	71	5691 / 1NY4	0.10	0.25	0.15
Non-structural protein 1 (NS1)[108]	73	66	4317 / 1NS1	0.03	0.41	0.38
Ribonuclease pancreatic (RnaseC6572S)[109]	124	116	4032 / 1SRN	0.42	0.20	0.22
Ribonuclease pancreatic (RnaseWT)[109]	124	116	4031 / 1SRN	-0.18	-0.25	0.07
Z domain of staphylococcal protein A [110]	71	67	5656 / 1H0T	2.75	2.69	0.06
Staphylococcus aureus protein SAV1430 (Target ZR18)[111]	91	85	5844 / 1PQX	-0.14	0.00	0.14

5.4 Discussion

5.4.1 Expectations and limitations of the statistical modeling

The underlying statistical modeling implemented in BaMORC also assumes that the ^{13}C chemical shifts approximately follow sets of standard distributions. Therefore, the best results are expected when the ^{13}C chemical shifts of each amino acid in any secondary structure follow a bivariate normal distribution with no overlap between distributions. We performed four goodness-of-fit tests for normality on each chemical shift distribution, which indicated that each distribution was approximately normal and reasonable to be used for parametric statistical purposes in our analysis; however, there is clear overlap between many of these distributions (Figure 3.5 and Figure 3.7). To ameliorate the distribution overlap status quo, we constructed a prediction overlap matrix and predictor weights using a Bayesian-inspired, reverse-logic approach. In addition, amino acid cysteine chemical shift data were classified into two unique distributions to minimize their overlap with other amino acid statistical models, which is justified by the presence of two oxidative states for cysteine residues in the normal cellular environment.

5.4.2 Bias correction and parameter optimization

During the development of the BaMORC methodology, we addressed several issues regarding chemical shift data quality in the RefDB entries, which are derived from the BMRB. The reference correction of BMRB protein entries provided by the RefDB was a starting point that enabled the derivation of amino acid and secondary-structure-specific expected values and variances for C_α and C_β resonances. However, we first had to split cysteines into two separate oxidative groups because of overlap problems created by the

wide cysteine distributions. Next, problems in inter-spectral registration decouple the assigned chemical shifts reported in the BMRB entries, which are passed onto the RefDB entries utilized in this work. Therefore, we developed several refinements of the RefDB to derive more accurate covariances, improving the performance of BaMORC. To further refine the covariance values, we filtered out all of the datasets that are likely not to come from a single NMR experiment. In the data filtration pipeline described in the Methods, we compared C_α versus C_β RMSD values of individual RefDB entries. The aim was to use only the entries that represented C_α and C_β shifts with strong covariance (e.g. derived from single experiments). Among 1557 entries, the correlation optimization filtered down to 729 entries for calculating optimal covariances. The resulting improvement between the inaccurate covariances and the optimal covariances is illustrated in Figure 3.12. Nearly all of the 60 covariances are improved, with some showing significant changes including a change in sign. These improvements, as visually illustrated in Figure 3.7e, demonstrate the improved accuracy of the resulting statistical models to represent the underlying NMR chemical shift data. Moreover, additional distinct distributions do appear present in Figure 3.7 and are due to the presence of other secondary structures and structural phenomena. For instance, it is well-known that cis/trans isomerization of proline has certain effect on secondary structure and affected chemical shift distributions[112]. These unaccounted chemical shift distributions can lower calculated covariance values. However, as more BMRB entries include 13C-assigned peak lists, we see an opportunity to further refine covariance statistics. According to our estimates, about $60\,(20\,amino\,acids \times 3\,secondary\,structures) \times 60 \times 50\,(minimal\,sample\,size) = 180{,}000$ ^{13}C-

assigned peaks are required in the BMRB for the next generation of covariance analysis. Currently, the BMRB contains approximately 11,500 ^{13}C-assigned peaks.

5.4.3 Reference correction performance on real data

We have tested the performance of the general BaMORC method in detecting reference correction values under various conditions. The reference correction values were within +/-0.45 ppm of the SHIFTX determined references at the 90% IQR with an absolute length of 0.73 ppm for datasets derived from the RefDB. The typical NMR dataset includes approximately 85% of the expected spin systems. Therefore, we tested our algorithm on incomplete data by incrementally removing a certain percentage of the data from each dataset tested. The robustness of the algorithm is stunning: it performs very well, maintaining referencing correction within [-0.41, 0.57] ppm range of the correct value at the 90% confidence level, even when 50% of the data are randomly removed. This robustness is achieved because the algorithm uses a non-parametric approach (i.e. a comparison of expected and predicted amino acid frequencies). Additionally, keeping reference correction within +/-0.6 ppm of the correct value is very important for accurate amino acid typing used in protein resonance assignment analysis and for accurate secondary structure analysis from chemical shifts. When carbon chemical shift referencing accuracy is outside the [-0.43, 0.64] ppm range, the relative error rate in amino acid and secondary structure prediction increases dramatically as illustrated by the increase in residuals in Figure 5.14.

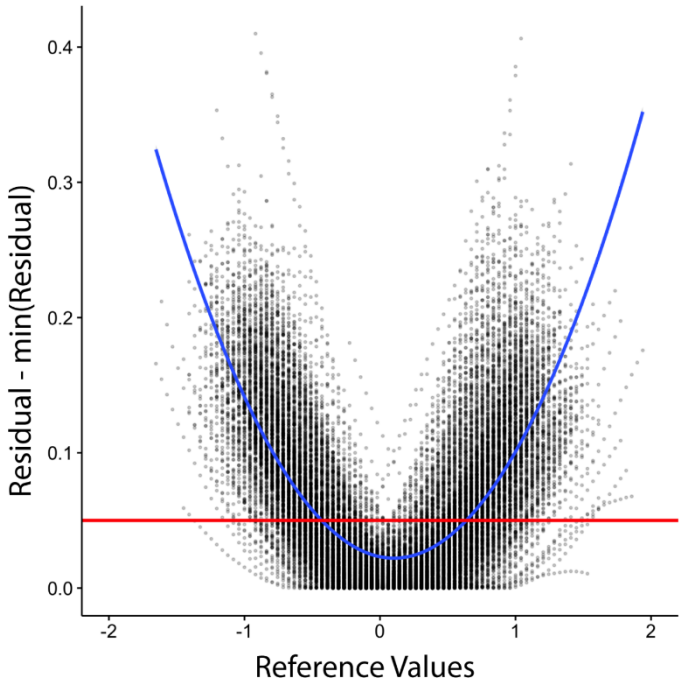

Figure 5.14 Amino Acid and Secondary Structure Frequency of Residual vs. Reference Correction Values for RefDB datasets. The y-axis is the residual of observed and predicted AA-SS frequencies from BaMORC minus the minimum residual observed corresponding to the Reference Correction Value on the x-axis. The blue line is the quadratic regression line to the values. The red line represents a 5% error rate above the best amino acid and secondary structure prediction performance. The intersection of the red line with the blue line occurs at -0.43 ppm and 0.64 ppm.

Also, this performance on spin system datasets derived from the RefDB completely translates to the real-world use-case where real, unassigned, experimental HN(CO)CACB peak lists are utilized. All peak list data were manually peak-picked. There are extra peaks in the data, which could be artifacts or from additional resonances due to multiple local protein conformations. Table 1 illustrates even better performance by Unassigned BaMORC on experimental peak lists, keeping chemical shift referencing within +/- 0.4 ppm for all 10 peak lists tested. While the sample size is small, i.e. only 10 experimental peak lists, the best Unassigned BaMORC performance may inferior to the Assigned BaMORC performance, which reflect the fact that many RefDB derived spin system

datasets come from multiple NMR spectra, weakening C_α/C_β correlation. Also, two of the experimental peak lists had a carbon chemical shift reference deviation that was over 2 ppm. Peak lists with large chemical shift referencing errors is the exact situation that Unassigned BaMORC was designed to detect and correct, so that a scientist does not waste time and effort trying to utilize such highly miss-referenced peak lists for downstream analyses, especially protein resonance assignment. The resulting assignments would be error prone and their chemical shifts would propagate error during structure determination. But even more subtle deviations in the 0.6 to 2.0 ppm range can have a significant impact on assignment and structural error. But Unassigned BaMORC has a demonstrated performance in keeping carbon chemical shift referencing within the +/- 0.4 ppm range.

5.4.4 Computational considerations

To evaluate the computational running time of the BaMORC algorithm, we measured execution time of calculations using the R function "system.time()" on the same computer system with the following specifications: CPU model Intel(R) Core(TM) i7-4930K CPU @ 3.40GHz with 6 CPU cores (12 with hyperthreading), Fedora 22 x86_64 operating system, and 64GB RAM. We tested both a version that utilized a grid-search approach for optimization and a version that utilized the Development Evolution optimization library (DEoptim) [101] for optimization. While these implementations are not parallelized, we did test 4 datasets at a time in 4 separate processes that each utilized a single CPU core. At the early stage of the research and development, we used a grid-search[113] as the optimization algorithm. Later we switched to using the DEoptim as a replacement for the grid-search approach.

The grid-search implementation evenly stepped across a range between -5 and 5 ppm in a first iteration and then evenly stepped across a +/- 1 ppm bounded minimum in a second iteration. In our testing, we used either 25 or 50 steps in each iteration.

As mentioned earlier, the datasets are from the RefDB and each dataset have a different number of chemical shift pairs. As show in Figure 5.15, the distribution of dataset sizes is centered around 100. The running time analyses were performed on a range of 50 to 150 chemical shifts pairs incremented by 10. To prepare the datasets for testing, we started with the 114 datasets out of 1557 total datasets with missing values removed that had at least 150 chemical shifts pairs per dataset. Then using these 114 datasets, we trimmed each dataset to the appropriate size for each increment: 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, and 150 chemical shift pairs, resulting in $11\ increments \times 114\ datasets = 1254\ datasets$ for running time measurements.

Each input file was then scanned and reformatted into tabular format. Finally all of the 1254 files were concatenated into a single R-object that is loaded prior to testing.

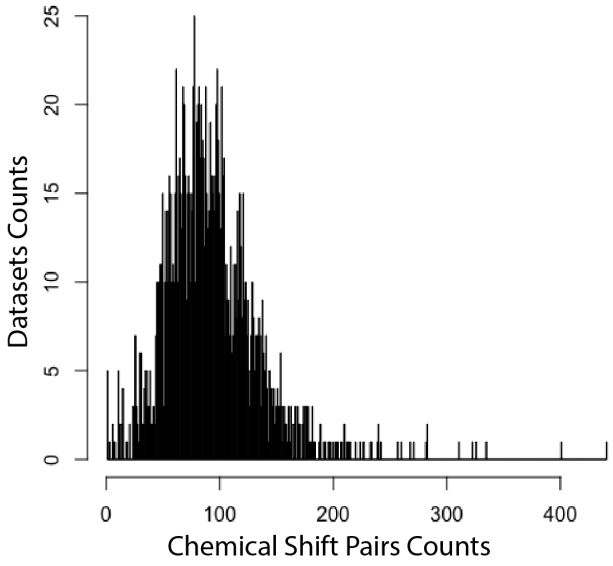

Figure 5.15 Datasets counts distribution based on the number of chemical shift pairs.

The BaMORC core function call calculates up to 60 probabilities for each chemical shift pair with respect to 20 amino acids (without glycine, but with cysteine in separate oxidized and reduced states) across three secondary structure types that are present in a given dataset. Due to resonance types (C_α and C_β) not being absolutely discernable from the grouping of experimental peaks into spin systems, each chemical shift pair must be evaluated twice as (C_α, C_β) and (C_β, C_α), doubling the number of probabilities calculated to a maximum of 120. Since the probability calculations dominated the operations in our algorithm, we focused on them. Therefore, for a given protein dataset with n chemical shift pairs, 120n probabilities are calculated at any given reference correction value, representing $O(n)$ complexity. With a grid-search algorithm, the complexity becomes $O(nm)$, where m is the number of reference correction grid points tested. We used two rounds of grid search at different granulations, each including 50 steps, for a total of 100 steps. To have a base line comparison, we also used the grid-search with 25 steps per

round, and the results are shown in Figure 5.16. As we expected, the computation time increases linearly as the number of chemical shift pairs increases, and the 100-step grid-search is double the 50-step grid-search. This further verifies the asymptotic $O(nm)$ complexity. Since different datasets varied in their completeness (i.e. number of chemical shift pairs present), a typical 10KDa (~100 residues) protein dataset which may have anywhere from 75% to 100% completeness is expected to take between ~7-10 seconds on the computer system described above, based on a 100-step grid-search.

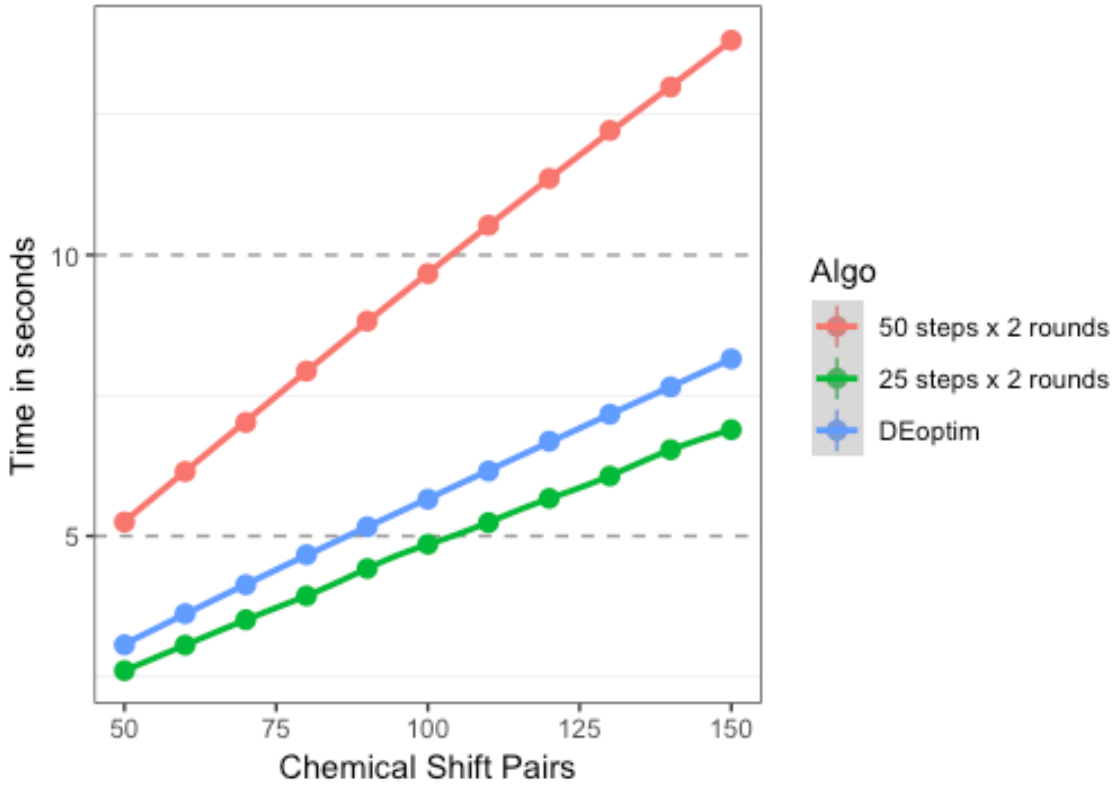

Figure 5.16 Execution time for the algorithm. Red: using two rounds of grid-searches with 50 steps; green: using two rounds of grid-searches with 25 steps; blue: using the DEoptim algorithm with max iteration set as 10. The results show that the execute time of all three algorithms increase in linear fashion as the dataset size grows.

To improve the computational timing and performance, we replaced the grid-search algorithm with a global optimization algorithm Differential Evolution (DE)

algorithm from the DEoptim[100]. The computation time increases linearly as the number of chemical shift pairs increases as shown in Figure 5.16. The DE algorithm was invented by R. Storn and K. Prince in 1997[102], and is an powerful, derivative-free global optimization algorithm that performs the optimization via the evolution of a population of candidate solutions. The core of the algorithm uses a process of evolution and belongs to genetic algorithms that uses biology-inspired operations of crossover, mutation, and selection over a population of potential solutions to optimize the objective function. The algorithm iteratively tries to improve candidate solutions, where the fittest individual solution of a population will produce more "offspring" solution that inherit the good traits, thus evolving the population of candidate solutions. One major advantage of this algorithm is that it has no requirement for the objective function to be differentiable. In fact, almost no restricting assumptions are required in contrast to many other optimization algorithms. The implementation of the algorithm[114] and the proof of its convergent properties are out of the scope per this dissertation. Since the DE algorithm uses a sampling approach, the run-time is determined by the number of iterations or the cut-off of the stop step. Although the DE algorithm doesn't guarantee that the returned value is the actual minimal value due to the nature of the non-deterministic approach; however, in our case, the DE algorithm provides reference correction values (returned minimum) that were at least equivalent (and generally superior) to the grid-search results with less running time. Concluding, 10 DE algorithm iterations generally provided an equivalent reference value within a -5.00 ppm to +5.00 ppm testing range, which is roughly two times faster than the grid-search approach we had previously employed with two rounds of 50 steps.

5.4.5 Model assumptions for appropriate use

An issue facing any model-based approach to data analysis is the validity of the model assumptions. The most important model assumptions here are that each pair of C_α and C_β chemical shifts is identical and independent, following a bivariate normal distribution, and the shapes of the distribution are well-represented by ellipses. Although we expect the algorithm to be robust to morphologically similar distributions, such as flat-top clusters or low-aspect-ratio ellipses, the algorithm is certainly not designed for the analysis of very small proteins or peptides. In addition, the presence of paramagnetic compounds, ring current effects, and deuteration shift effects will generate outlier chemical shift values that significantly deviate from the expected values derived from the RefDB dataset.

The default assumptions stipulate that each input dataset is at least 50% complete, meaning that the number of missing spin systems should not represent more than 50% of the expected number of spin systems based on the protein sequence. In practice, we found datasets with greater than 70% completion produced consistent reference correction values. If the user wishes to statistically demonstrate the applicability of our approach to a problem, they can employ the residual (sum of the absolute difference) plot. We have thoroughly tested our defaults assumptions on a wide variety of protein scenarios (e.g. all of the relevant entries in the RefDB) and found the correction results to be largely insensitive to protein classification. However, we recognize that there are extreme examples like disordered proteins for which these choices may not be advised. As with all Bayesian analyses, it should be remembered that the prior parameters should genuinely represent the subjective prior beliefs.

5.4.6 Pragmatic implementation decisions and future development

Unassigned BaMORC is currently designed to correct ^{13}C chemical shift referencing using HN(CO)CACB-type peak lists. The focus on ^{13}C chemical shift referencing is pragmatic from three perspectives: i) C_α and C_β provide the most information about amino acid type, which is central to the BaMORC methodology; ii) accurate ^{13}C chemical shifts have the greatest impact on protein resonance assignment and other downstream analyses; and iii) grouping of the HN(CO)CACB peaks into spin systems is more robust than for other NMR experiments. Likewise, Assigned BaMORC is designed to use assigned C_α and C_β chemical shifts for reference correction after initial chemical shift assignment, but before other downstream analyses. However, we are pursuing further improvements to the methodology and current implementations. We see a host of possible improvements that would extend the methodology to correct ^{1}H and ^{15}N chemical shift referencing and allow the application of the method to peak lists derived from other types of NMR experiments as well. Though, some of the improvements will require further evaluation and refinement of the chemical shifts from BMRB and RefDB entries and may require waiting until sufficient assigned peak lists are present in these public scientific repositories. For instance, developing an extension to handle intrinsically disordered proteins (IDPs) would likely require more than the 176 IDP BMRB entries available as of May 2018.

5.5 Conclusions

The BaMORC method utilizes unassigned C_α and C_β chemical shift data to generate accurate ^{13}C reference correction within +/- 0.45 ppm at the 90% confidence level on

RefDB derived test datasets. BaMORC also demonstrates robust performance, keeping the ^{13}C reference correction within +/- 0.6 ppm at the 90% confidence level even with up to 50% of the ^{13}C chemical shift data missing. Keeping the reference correction within 0.6 ppm of the correct value is very important for accurate amino acid typing to be used in protein resonance assignment analysis. The Unassigned BaMORC method utilizes unassigned C_α and C_β chemical shift data from HN(CO)CACB-type experimental peak lists to generate accurate ^{13}C referencing correction within +/- 0.4 ppm for all 10 HN(CO)CACB-type experimental peak lists tested. The Assigned BaMORC method utilizes assigned C_α and C_β chemical shift data to generate accurate ^{13}C chemical shift reference correction within +/- 0.22 ppm at a 90% confidence interval. Unassigned BaMORC can correct ^{13}C chemical shift referencing at the beginning of protein NMR analysis, when accurate ^{13}C chemical shift referencing is needed the most for accurate protein resonance assignment, structure determination, and other downstream analyses. Assigned BaMORC can refine the referencing once assignments are made. Additionally, the underlying BaMORC method is robust to missing ^{13}C chemical shift data, which addresses the real-world situation of incomplete ^{13}C resonance detection. Therefore, the BaMORC methods will allow non-NMR experts to detect and correct ^{13}C referencing error at critical early data analysis steps, lowering the bar of NMR expertise required for effective protein NMR analysis.

CHAPTER 6. BAMORC PACKAGE FOR ACCURATE AND ROBUST ^{13}C REFERENCE CORRECTION OF PROTEIN NMR SPECTRA

6.1 Overview

BaMORC, a statistical software package that performs ^{13}C chemical shifts reference correction for either assigned or unassigned peaks lists derived from protein NMR spectra. BaMORC provides an intuitive command line interface that allows non-NMR experts to detect and correct ^{13}C chemical shift referencing errors of unassigned peak lists at the very beginning of NMR data analysis, further lowering the bar of expertise required for effective protein NMR analysis. Furthermore, BaMORC provides an application programming interface for integration into sophisticated protein NMR data analysis pipelines, both before and after the protein resonance assignment step.

6.2 Introduction

Chemical shifts derived from protein NMR spectra have a wide variety of uses including protein structure determination[25,26], characterizing ligand binding[115-117], and drug discovery and design[60,66]. However, deriving accurate chemical shifts values requires the referencing of NMR spectra to a certain standard, typically an internal standard[69,118]. Due to human errors and variety of experimental factors[70,119], variance, or errors occur quite frequently in 13C protein NMR data. An estimated 40% of the entries in the Biological Magnetic Resonance Bank (BMRB) have referencing issues[120]. The resulting referencing discrepancies are highly problematic since prior methods for reference correction required either assignment and/or structure[71,97], which are the exact downstream aims that reference correction is trying to target. This leads to a co-dependency between

reference correction and NMR structure determination, crippling the progress of many protein NMR downstream analyses[116,121-125].

We therefore developed the Bayesian Model Optimized Reference Correction (BaMORC) method [126] that helps non-expert scientists to detect and correct C_α and C_β chemical shifts, at the beginning of the protein NMR analysis process, when chemical shifts are unassigned. Here we describe the BaMORC method implemented in an easy-to-use software package written in the R programming language. BaMORC uses a Bayesian model to estimate an amino acid frequency from C_α and C_β chemical shift statistics inferred from the Re-referenced Protein Chemical shift Database (RefDB) [86], with or without resonance assignment information. As shown in Figure 4.1, by optimizing the minimal between the actual amino acid frequency calculated from known protein sequence and an estimation based on the observed chemical shifts, BaMORC returns the reference correction value and re-referenced chemical shifts data. Figure 6.1 illustrates the required input and expected output generated by the BaMORC R package.

6.3 Overview of the BaMORC package

The BaMORC R package provides a command-line interface (CLI) for general use and an application programming interface for users that are familiar with R programming, especially for use within an integrated development environment like RStudio [127]. As illustrated in Figure 6.1, the BaMORC R package can use the protein sequence and chemical shifts in a variety of unassigned and assigned formats including the NMR-STAR format utilized by the BMRB. The general row-based text format may be delimited by comma or white space, but with the protein sequence on the first line followed by unassigned peaks or assigned C_α and C_β chemical shift pairs on following rows.

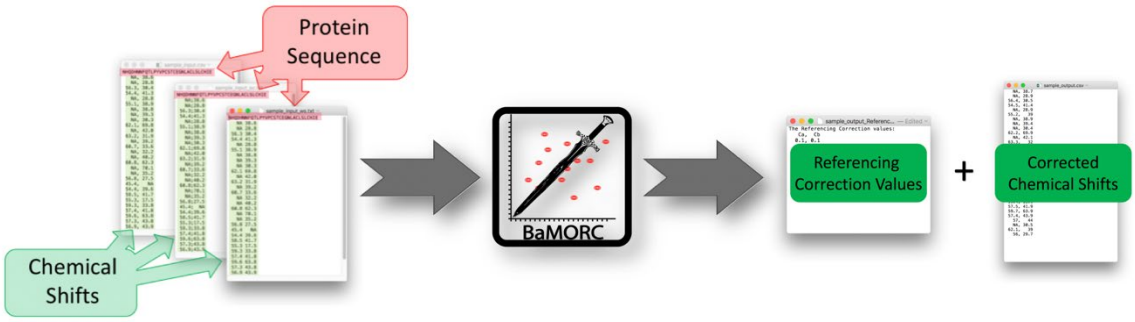

Figure 6.1 Required input and expected output of BaMORC R package.

Each input file is referred to as a "task" within a larger "job". The BaMORC R package automatically interfaces with the registration, grouping and referencing algorithms to set up tasks and print out most optimized correction values for a give input, and returns the corrected chemical shifts in csv format. The package can also accept a BMRB ID such as BMR 4020 as input to retrieve corresponding files from the BMRB web server, automatically parsing the file, correcting the referencing, and returning the same set of output as mentioned before.

We have evaluated BaMORC against 568 ^{13}C protein NMR datasets from the RefDB with 90% or higher completeness with respect to C_α and C_β chemical shift assignments. Outputted reference correction values should match closely to 0 ppm, since each dataset from RefDB has been reference corrected using protein structure information. With chemical shift assignments, BaMORC provides reference correction values within +/- 0.50 ppm for all datasets and within +/- 0.22 ppm for 90% of the datasets, representing a 90% Confidence Interval (CI) of 0.40 ppm (see Figure 5.11)[126]. This level of performance is superior to the prior state of the art LACS method[51].

However in the real-world situation, ^{13}C reference correction is most valuable before protein resonance assignments are known. This situation is what the BaMORC

package was really designed to address. The unassigned BaMORC method has two major components, grouping and referencing correction. With an input peak list, the grouping algorithm will return a list of C_α and C_β grouped peaks (spin systems) as output, which will be the input for the referencing correction algorithm as shown in Figure 6.1. The grouping algorithm is a variance-informed DBSCAN algorithm that employs derived dimensions-specific match tolerances values to group peaks into spin systems. A peak list registration step is used to derive the necessary match tolerance values [16]. In addition to the grouped peaks, the referencing correction component uses the JPred4 [17] server to generate sequence-based secondary structure predictions and then calculates the reference correction.

Again, we used the same 568 ^{13}C protein NMR datasets from the RefDB to evaluate the reference correction component of Unassigned BaMORC, but without chemical shift assignments. As shown in Figure 4, the reference correction component of Unassigned BaMORC provides reference correction values within +/- 0.45 ppm for 90% of the datasets, representing a 90% CI of 0.69 ppm[126]. This suggests that the unassigned BaMORC algorithm can achieve the same level of performance when handling unassigned ^{13}C protein NMR peak list data. This level of real-world performance is demonstrated with a set of peak lists derived from solution NMR HN(CO)CACB spectra for 10 different proteins. In this real-world evaluation, Unassigned BaMORC provided reference correction values all within +/- 0.40 ppm[126].

6.4 Materials and Methods

6.4.1 Software

The Python programing language, version 3.6, is used for the grouping algorithm. The R programming language, version 3.4, is used for the BaMORC core component. The library dependencies are listed below:

Python Library Dependencies: Python (>=3.6), gcc (>=5.1)

R Library Dependencies: R (>=3.4), data.table, tidyr, DEoptim, httr, docopt, stringr, jsonlite, readr, devtools, RBMRB, BMRBr

6.4.2 Experimental data sources

All the data are from the RefDB are used to derive chemical shifts statistics within the BaMORC package. For testing and evaluation, we used datasets from the RefDB and experimental peaks lists from a variety of sources.

6.5 Installation

To use the BaMORC package, users must first install the R 3.4.x (or higher version) and Python 3.6.x (or higher version) interpreters on their machine. For Linux distributions, this is typically accomplished through the distribution's package management system. For other operating systems, installation may require a more manual procedure. R language is a language and environment for statistical computing [128]. The installation guide is located in the website of the comprehensive R Archive Network [https://cran.r-project.org/]. Python language [129] can be install from this website [https://www.python.org/].

6.5.1 Install from command line (Linux and Mac only)

To use BaMORC, the user first needs to install the package from the GitHub or CRAN.

```
$ wget -q https://cran.r-project.org/src/contrib/BaMORC_<version>.tar.gz
$ sudo R CMD INSTALL BaMORC_<version>.tar.gz
```

6.5.2 Install from command line via R console

```
$ R # to start R console
> install.packages("BaMORC")
```

6.5.3 Install from R console

```
> install.packages("BaMORC")
```

6.5.4 Installing unassigned BaMORC dependencies

The unassigned BaMORC analysis requires the ssc (Spin System Creator) package, which includes a variance-informed implementation of the DBSCAN algorithm used for protein NMR spin system clustering. A docker container including the ssc package is required. Therefore, the user needs to install both docker and SSC docker image.

- Install Docker from https://www.docker.com/products/docker-desktop.
- Install SSC docker container after docker is installed by running following code:

```
> docker pull moseleybioinformaticslab/ssc.
```

6.6 The BaMORC application programming interface (API)

After import the BaMORC in R either on R Console or in RStudio, the user will first read in NMR chemical shifts data via the read_file function with parameters of file path, file delimiter, and a flag that indicates whether data is assigned or unassigned. BaMORC currently support file delimiters of comma, semicolon and whitespace. For users who want to run an analysis on an existing dataset from the BMRB (NMR-STAR version 2 and 3), they can use either the read_nmrstar_file function with a parameter for a local file path or the read_db_file function with a parameter for the BMRB ID and a flag that indicates whether data is assigned or unassigned. If read_db_file is used, BaMORC will utilize the BMRB web API to fetch the corresponding BMRB entry matching the ID. Table 6.1shows common usage patterns for reading input data into the BaMORC referencing correction analysis pipeline. For a full list of available conversion options and more detailed examples and documentation of all the functions, please refer to “The BaMORC Reference” and “Quickstart.”

Table 6.1 Summary of BaMORC package interface (API)

Command	Description	Example
read_file	Import local files	input_data = read_file(file_path = "./sample_input.txt", delim = "ws", assigned = T)
read_nmrstar_file	Import files in NMR-STAR format	input_data = read_nmrstar_file ("BMR4020.str")
read_db_file	Use BMRB ID to import files	input_data = read_db_file(id = "BMR4020")
bamorc	Using sequence, secondary structure and chemical shift data to estimate the reference correction value	bamorc(sequence, secondary_structure, chemical_shifts_input, from=-5, to=5)
unassigned_bamorc	Using only sequence and chemical shift data to estimate the reference correction value	Unassigned_bamorc(sequence, chemical_shifts_input, from=-5, to=5)

Next, the user will pass the input data as parameters to the bamorc() or unassigned_bamorc()function, which will perform the reference correction analysis. Both functions utilize the output from the read-in functions mentioned above and will perform a secondary structure estimation based on the provided protein sequence if secondary structure information is not provided. Through a series of optimization calculations (details refer to paper [126]), bamorc() and unassigned_bamorc() will return the estimated referencing correction value in a plain text file and corrected chemical shifts for both Cα and Cβ as a table as shown in Figure 6.1. The user can optionally customize the search range. Table 1 contains a basic example of calling each function. For detailed examples and expected outputs of BaMORC API functions, please refer to the online documentation: https://moseleybioinformaticslab.github.io/BaMORC/index.htm.

6.7 The BaMORC Command Line Interface (CLI)

The BaMORC CLI is an extension of the BaMORC package, aimed at the broader NMR community that is not familiar with R programming language. After installing the BaMORC package and the ssc Docker container (if for unassigned protein NMR analysis) as shown in the installation section. To use BaMORC CLI, the user needs to find the CLI run-script first by opening a terminal and typing the command highlighted in Figure 5.

```
> R -e 'system.file("exec", "bamorc.R", package = "BaMORC")'
```

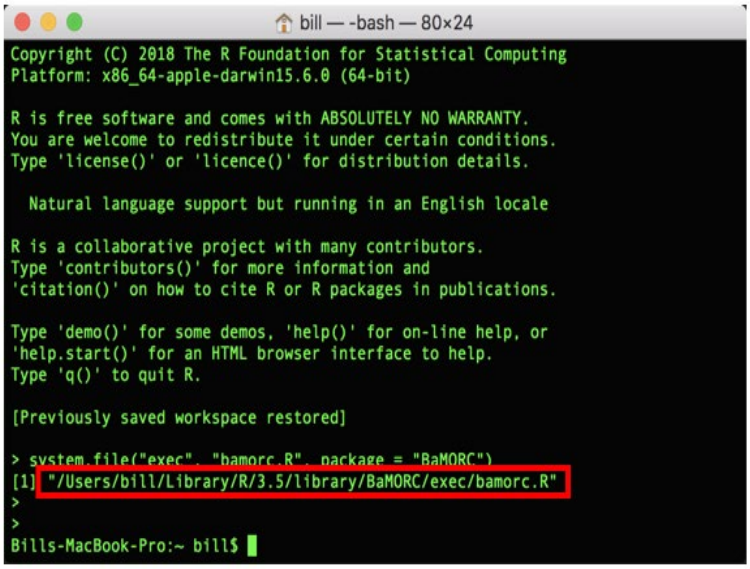

Figure 6.2 Finding the CLI run-script location.

The user can then execute the appropriate command listed in Table 2 to run an analysis. Similar to the package, the BaMORC CLI has three major modules: assigned and unassigned reference correction for assigned and unassigned protein NMR data and a

miscellaneous collection of other useful tasks. Table 6.2 list the components of the CLI and their associated parameters.

Table 6.2 BaMORC CLI commands and their parameters.

Command	Parameter	Example
Assigned	Required parameter:	
	Input file path or ID	--table=sample_input.csv or --bmrb=bmr4020 or --id=BMR4020
	Optional parameter:	
	Estimation range	--range=(-5,5)
	Delimiter	--delim=comma
	Output path	--output=sample_output.csv
	Report file path	--report=sample_report.txt
Unassigned	Required parameter:	
	Input file path	--table=sample_input.csv
	Optional parameter:	
	Grouped peaklist or not	--grouped=true
	Protein sequence	--seq=sample_sequence.txt
	Search range	--range=(-5,5)
	Output path	--output=sample_output.csv
	Report file path	--report=sample_report.txt
Help	Help menu	--h or -help
Version	Version number	--v or -version

To help the user transition between the API and CLI, Table 6.3 illustrates common BaMORC CLI usage examples with corresponding BaMORC API examples. The CLI is utilized within a command line terminal on Linux and Mac computers. For windows user,

please refer to our online documentation for more details. We have developed online documentations, available at:

(https://moseleybioinformaticslab.github.io/BaMORC/index.html).

Table 6.3 BaMORC CLI usage and corresponding API commands.

CLI	API
Assigned BaMORC: For user's own protein NMR spectra result	
$ bamorc.R assigned --table=./sample_input.csv --ppm_range=(-5,5) --output=./sample_output.csv --delimiter=comma --report=./sample_report.txt	> user_input = read_file(file_path="./sample_input.csv", delim="comma", assigned=f) > result = bamorc(sequence = user_input[[1]], chemical_shifts_input = user_input[[2]], from = -5, to = 5)
Assigned BaMORC: For data in NMR-STAR format	
bamorc.R assigned --bmrb=BMR4020.str --ppm_range=(-5,5) --output=./sample_output.csv --delimiter=comma --report=./sample_report.txt	> bmrb_format_data = read_nmrstar_file("BMR4020.str") > result = bamorc(sequence = bmrb_format_data[[1]], chemical_shifts_input = bmrb_format_data [[2]], from = -5, to = 5)
Assigned BaMORC: For data already existing in BMRB database	
bamorc.R assigned --id=BMR4020 --ppm_range=(-5,5) --output=./smple_output.csv --delimiter=comma --report=./sample_report.txt	> existing_data = read_db_file(id="BMR4020") > result = bamorc(sequence = existing_data[[1]], chemical_shifts_input = existing_data [[2]], from=-5, to=5)
Unassigned BaMORC: For user's own protein NMR spectra result	
bamorc.R unassigned table=./sample_input.csv --ppm_range=(-5,5) --output=./sample_output.csv --	> user_input = read_file(file_path="./sample_input.csv", delim="comma")

delimiter=comma --report=./sample_report.txt	> result = unassigned_bamorc(sequence = user_input[[1]], from = -5, to = 5)
BaMORC CLI: other commands (CLI only)	
bamorc.R valid_ids	To show all the valid BMRB file IDS
bamorc.R -h	To show help menu
bamorc.R -v	To show BaMORC version

6.8 Conclusions

The BaMORC package is a useful R package, providing referencing correction for assigned and unassigned protein NMR data alone with several data parsing, data processing and calculation functions. Also, BaMORC provides a simple command-line interface that allow a broader usage in the NMR data center for reference correction and validation. Further information on the algorithms mentioned above and their development is available on the repository such as CRAN and GitHub. And source code is available at https://github.com/MoseleyBioinformaticsLab/BaMORC. The package has been submitted to CRAN and should be available from CRAN soon. We will add a sentence about its availability from CRAN and update installation instructions when the evaluation process is finished. The code is published under a modified open source BSD-3 license. Academic researchers are free to use it without restriction, except for proper citation. This repository includes code for the BaMORC referencing correction pipeline. For the registration and grouping algorithm, please refer to https://github.com/MoseleyBioinformaticsLab/ssc[28]. For further information and assistance please visit our laboratory website: http://bioinformatics.cesb.uky.edu.

CHAPTER 7. BAMORC WEB APPLICATION FOR STREAMLINE PREPROCESSING PROTEIN NMR SPECTRA

7.1 Overview

Procedures for preprocessing protein nuclear magnetic resonance (NMR) spectra involves numerous steps to properly transform and then reference the chemical shift data before later analyses. With respect to referencing, researchers either create post hoc workflows for each spectrum using existing protein structure or coordination information, or ad hoc preprocessing workflows by using internal referencing scheme. Over time, the complexity of these workflows has grown to handle various sample-dependent issues that can hamper the referencing accuracy, partially driven by the increased use of chemical shifts in protein structure determination. We introduce Bayesian Model Optimized Reference Correction (BaMORC), a method, software package, and web-based application to help address the challenge of robust and accurate referencing of NMR spectra. BaMORC adopts a streamlined preprocessing workflow to correct the ^{13}C referencing of protein NMR spectra both before and after the resonance assignment step, producing a final reference correction within +/- 0.2 ppm (i.e. 0.4 ppm at the 90% confidence interval). By introducing a statistical Bayesian model into the referencing optimization, BaMORC enables ^{13}C reference correction of without utilizing any prior information such as structural or assignment information from secondary experiments or analyses. BaMORC equips researchers with an easy-to-use and transparent web-based application, which is part of the R-package suites including command line and application programming interfaces that can be inserted into any protein NMR preprocessing workflow, improving the reliability of the downstream protein NMR analysis results for researchers who are novice to NMR technology.

7.2 Introduction

NMR is a commonly used technique for studying protein structure and dynamics. However, due to the intrinsic properties of NMR experiments, output data from NMR instruments requires a referencing value to be usable for down-the-line analyses. Poor chemical shift referencing, especially for ^{13}C in protein NMR experiments, fundamentally limits and even prevents effective study of the molecule(s) of interest. The primary goal of NMR spectral preprocessing is to transform the raw, collected data into a spectral representation that is interpretable with respect to the structural and dynamic properties of the molecules in the sample from which it was collected. In particular, preprocessing should identify any referencing inaccuracies and reduce their effect on the resulting spectral data and downstream analyses starting with resonance assignment. Accurate referencing, especially ^{13}C referencing is fundamentally important to prevent chemical shift deviations and mis-assignment that can lead to unrealistic structural and dynamic representations of molecules of interest, in particular proteins. An example of false or mis-assignment, familiar to most researchers, is an amino acid spin system or chemical shift mapped to the wrong amino acid in the primary structure of a protein, often due to poorly-referenced chemical shift values[26,59]. Although pain-staking, hand-assignment and evaluation approaches performed by expert NMR spectroscopists can overcome these types of errors for ill-referenced spectra of small and medium-sized proteins. However, avoiding these errors is highly burdensome, time-consuming, and error-prone for novice experimentalists typically using automatic or computer-assisted assignment and structure prediction/elucidation approaches that are meant to prevent the introduction of human error [115]. Also, many examples of machine learning or artificial intelligent approaches for

(semi-)automated protein NMR resonance assignments has illustrated the importance of data preprocessing [116,121-125].

The NMR community is well-equipped with tools that perform evaluations on the quality of chemical shifts after assignment and structure determination, including AVS [66], PANAV [67], CheckShift [67,68], SHIFTX2 [69] and VASCO [70], to name a few. Despite the wealth of the accessible software and multiple attempts to outline best practices for preprocessing, the large variety of protocols has led to the use of post hoc pipelines that require results from downstream analyses or an external secondary experimental result such as a 3D structure

These issues in ^{13}C referencing and their effects on downstream data analyses provided the rationale for the development of Bayesian Model Optimized Referencing Correction (BaMORC) [61]. BaMORC represents both a methodology [61] and R package[130]that has an application programming interface (API), command line interface (CLI), and a new web-based graphical user interface (webGUI) that can be flexibly inserted into any protein NMR data processing workflow both before and after the resonance assignment step. Figure 4.1 highlights these differences between a traditional protein NMR data processing workflow and BaMORC-enabled data analysis workflows. The BaMORC R package is available in the Comprehensive R Archive Network (CRAN) [126] and on GitHub[130].

7.3 Methods

Data used to derive statistics for the BaMORC algorithm are from RefDB also available in the package as data frame. Datasets are originally from BMRB and later

included in the RefDB after being corrected and verified against 3D protein structure[69,71]. However, it is well recognized that NMR repository include many inaccuracies and errors, and the RefDB is no exception. Data representation is crucial for the accuracy and robustness of BaMORC algorithm, and many included datasets in RefDB were collected via a sequential manner, i.e. combine two different experiments. For instance, C_α and C_β chemical shifts could either be from the same experiment, for instance an HNcoCACB NMR experiment or two experiments, for instance HNcoCACB and HNcoCA NMR experiments (Figure 3.10). If C_α and C_β chemical shifts are reported from two separate experiments, the vital statistics covariance or joint variability can be lost, destroying the ability to accurately calculate the covariance from a dataset. Just as the requirement for many biological measurements, the chemical shifts for both alpha and beta carbons should be measured from the same experiment, i.e. measurable phenomenon. Data selection criteria were described in this paper[61].

In addition, data completion and data entry size were another two important perspectives that need to be considered. Almost all the datasets are incomplete and data sizes range from less than 50 to several hundred spin systems. To ensure generalization of the model, we included only comprehensive datasets from the same experiment for derivation of statistics, but for several validation of the performance and robustness, we included all of the available data with 50 or more spin systems and having both C_α and C_β chemical shifts.

The BaMORC implementation. BaMORC has been developed using R, while Python was used for implementing the grouping algorithm SSC, and R Shiny-generated HTML was used to implement the web graphic user interface. Currently, the BaMORC

web app counts two default reference correction procedures: one for assigned spectrum and another for unassigned. Both pipelines can operate without secondary structure as input, since a Rest API was used behind the scenes to fetch the secondary structure prediction information from JPred using the input protein sequence [97]. The BaMORC R package includes the main BaMORC API functions, utility functions, a straightforward command line interface (CLI), and the web app. The BaMORC GitHub repository provides the source code, detailed documentation and use-cases, and a tutorial that can be used by anyone with basic command line skills across major operation systems. The BaMORC web app as shown in Figure 4.1, can execute with two essential inputs, the protein sequence and chemical shift values for alpha and beta carbons, along with one optional input, secondary structure. Both the protein sequence and secondary structures are required to be in single-letter formats.

Intra-peaklist spin system registration and *grouping for unassigned NMR spectra.* For unassigned NMR spectra, the very first preprocessing step is to group peaks into spin systems so that each spin system will have multiple peaks from corresponding resonances. In the SSC algorithm[120], the registration step derives the necessary match tolerance values from the peak list and then group peaks, based on a variance-informed version of the density based clustering algorithm DBSCAN [103].

Reference correcting grouped peaklist. For both unassigned and assigned NMR spectra, the start input is grouped peaklist file. For unassigned spectrum especially, the BaMORC will automatically group them using SSC library and return the appropriate input. As for an assigned spectrum, the user provided input should already be grouped. The BaMORC web app accepts both copy-and-pasted input or direct uploaded input files.

Another essential input is the known protein sequence in single-letter format. Sequence information can also be copy-and-pasted or directly uploaded into the web app. The extra information about secondary structures, which is also required in single-letter format, is optional. BaMORC can use the JPred Rest API to fetch predicted secondary structure based on user-provided sequence information. However to suit users who already estimate this piece information or those who want to use a different secondary structure prediction algorithm than JPred, we leave this option open. Secondary structure information dramatically improves the power of Bayesian Statistical model, and it provides a more refined information that the optimization algorithm will use to find the best referencing value.

BaMORC is thoroughly documented, open-sourced, community-driven, and developed with high-standards of software engineering at heart. All the functions included in the BaMORC package are well-documented and the web app includes detailed instruction and examples. The open-source nature of BaMORC will permit more frequent code reviews and model assessments that effectively enhanced the software quality and reliability[131].

Ensuring reproducibility with strict versioning and containers. For enhanced reproducibility, BaMORC fully supports execution via the Docker[132] and Singularity container platforms[133]. Docker or singularity images are generated and uploaded to a public container repository for each new update of BaMORC. These container images are released with a set of software versions which also include the version of dependent libraries and OS system. This helps to maximize run-to-run reproducibility and to address

the widespread compatibility issue and results variabilities due to the lack of reporting software versions.

7.4 Results

BaMORC is a robust and convenient tool that enables researchers and novice to prepare protein NMR carbon datasets from both assigned and unassigned spectra for analysis (Figure 4.1 bottom). Its outputs allow for a range of applications, including structure determination, dynamics calculation, protein-protein interaction analysis, and others. In addition, BaMORC can be utilized as re-referencing tools for reference error correction.

7.4.1 A modular design alongside allows for a flexible, and adaptive workflow.

BaMORC is composed of six components (i.e. modules) that enable two default ^{13}C reference correction procedures that can handle unassigned peak lists or assigned carbon chemical shifts datasets (Figure 7.1). Several of the components rely on other open-source packages, such as SSC for spin system creation [134], BMRBr[135] for downloading BMRB entries [26], and the JPred server for secondary structure prediction [97]. In particular, the secondary structure prediction component allows BaMORC to predict the residue-specific secondary structure using the JPred prediction server given an input amino acid sequence. BaMORC's modular design implements several mechanisms for utilization including an application programming interface (API), command line interface (CLI), and a web-based graphical user interface. This combination of interfaces provides a flexibility for incorporating BaMORC into almost any protein NMR data analysis workflow or

pipeline, whether it be tight integration within an R program, part of an automated command line-driven pipeline, or part of a web-based data analysis workflow.

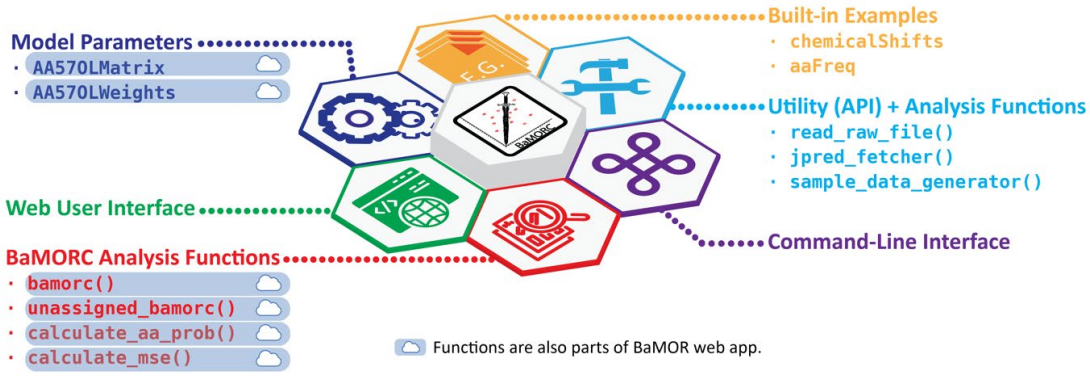

Figure 7.1 Modular design of the BaMORC package and web-based application. Six components comprise the package. Many components can operate independently, facilitating integration into other platforms and workflows.

7.4.2 BaMORC yields high-quality results, even from lower-quality datasets.

We have iteratively tested the robustness and overall quality of the results generated from BaMORC by using a three-stage validation approach (Figure 7.2). In stage one, we tested BaMORC using 568 entries collected from the Re-referenced Protein Chemical shift Database (RefDB). RefDB is a protein NMR data repository with all the chemical shifts from a Biological Magnetic Resonance Bank (BMRB) [26,71] entry carefully re-referenced using the SHIFTX-predicted chemical shifts based on corresponding 3D protein structures in the worldwide Protein Data Bank (wwPDB). Stage one concluded that BaMORC achieved a 0.7 ppm 90% confidence interval on unassigned CA/CB chemical shifts and a 0.4 ppm 90% confidence interval on assigned CA/CB chemical shifts [61]. We also compared assigned BaMORC's performance to the prior state of the art linear analysis of chemical shifts (LACS) method. As illustrated in Figure 7.3, assigned BaMORC dramatically outperforms LACS. In the stage two, we tested the resilience of the unassigned BaMORC

method against lower quality data such as datasets with missing values, which is typical for protein NMR datasets due to a variety of experimental issues. As illustrated in Figure 7.3, the overall performance of BaMORC doesn't deteriorate appreciably, even with only 70% of the ^{13}C spin systems present [61]. Stage three tested practical issues with using BaMORC in a real-world setting. These tests included comparing the use of predicted secondary structure using JPred versus secondary structure derived from the RefDB. As illustrated in Figure 7.2, there were almost no differences in the reference correction performance of the BaMORC algorithm. Also, the full unassigned BaMORC procedure was tested using a set of unassigned three-dimensional HN(CO)CACB-type experimental peak lists. All datasets were kept with +/- 0.4 ppm of the original expert-derived reference correction.

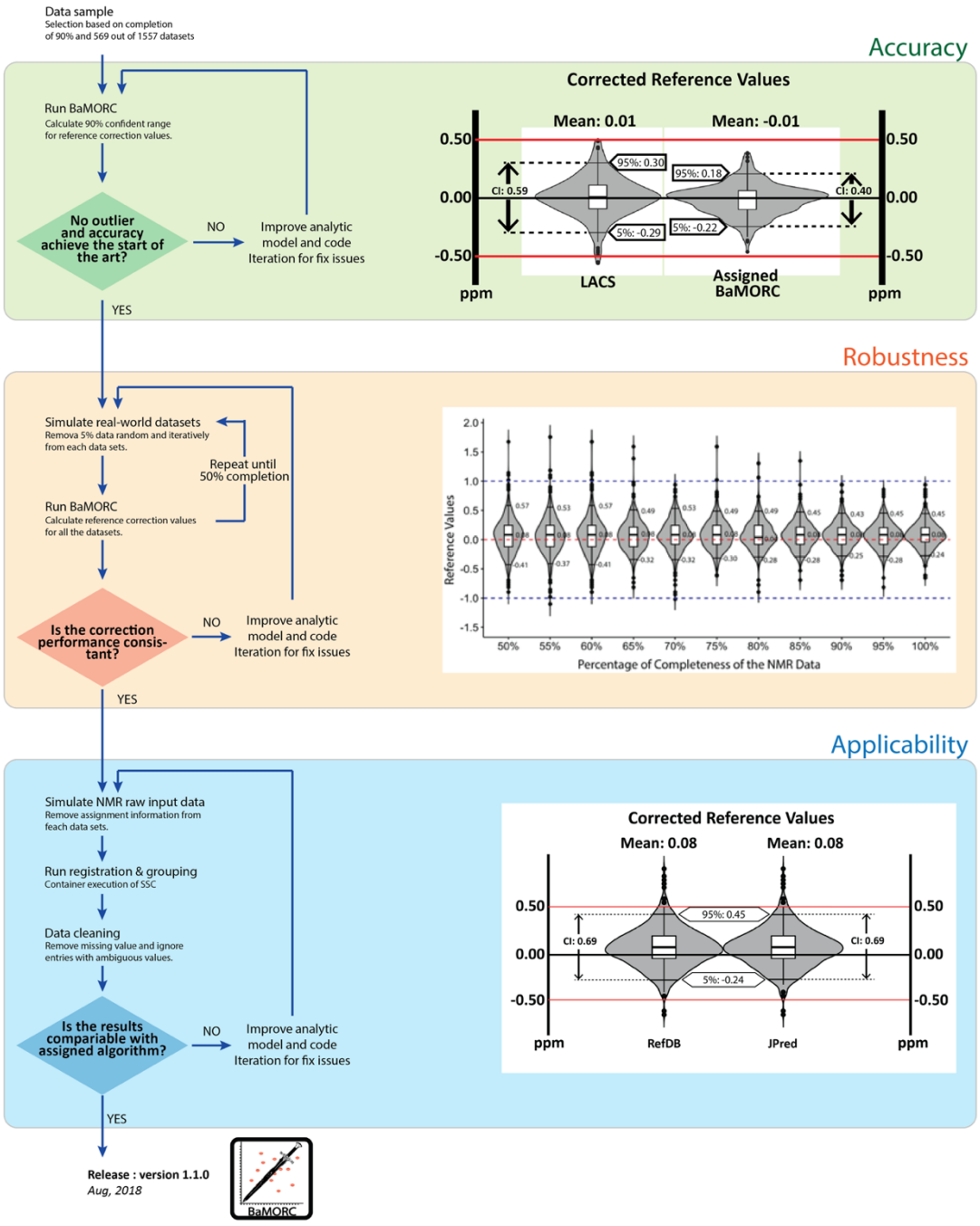

Figure 7.2 Three validation stages. Iteratively tested the robustness and overall quality of the results generated from BaMORC by using a three-stage validation approach: stage one tests the accuracy of the BaMORD; stage two tests the robustness; and stage three for the general applicability to real-world datasets.

7.4.3 Web-based graphic user interface with reporting functionality.

As illustrated in Figure 7.3, BaMORC's graphical user interface allows a user to correct ^{13}C referencing in a protein NMR dataset through an easy-to-use web browser-based interface that can be run in a standalone mode via the BaMORC R package. The web app has two major sub-interfaces tailored for either assigned and unassigned dataset analysis modes. The left side (panel) of Figure 7.3 shows the unassigned BaMORC sub-interface, which requires a protein sequence and a peaklist with C_α and C_β chemical shifts. The assigned BaMORC requires a protein sequence and assigned (amino acid typed) C_α and C_β chemical shift pairs. Optionally, residue-specific secondary structure can be provided in each sub-interface as well. The web app accepts direct copy-and-paste into the web interface or file upload inputs. In addition, users can also assess the quality of the ^{13}C reference correction with an individual report generated per protein dataset. The instructions are documented on the right panel of the interface as shown in Figure 7.3. The output reports are generated in HTML as shown in Figure 7.4, which can be opened with any web browser and contain key results: i) reference correction values from BaMORC algorithm and ii) the corrected output data either in grouped or ungrouped format. Additional csv and JSON output formats allow interoperability with other NMR data analysis tools, enabling the integration of BaMORC into web-based data analysis pipelines and workflows.

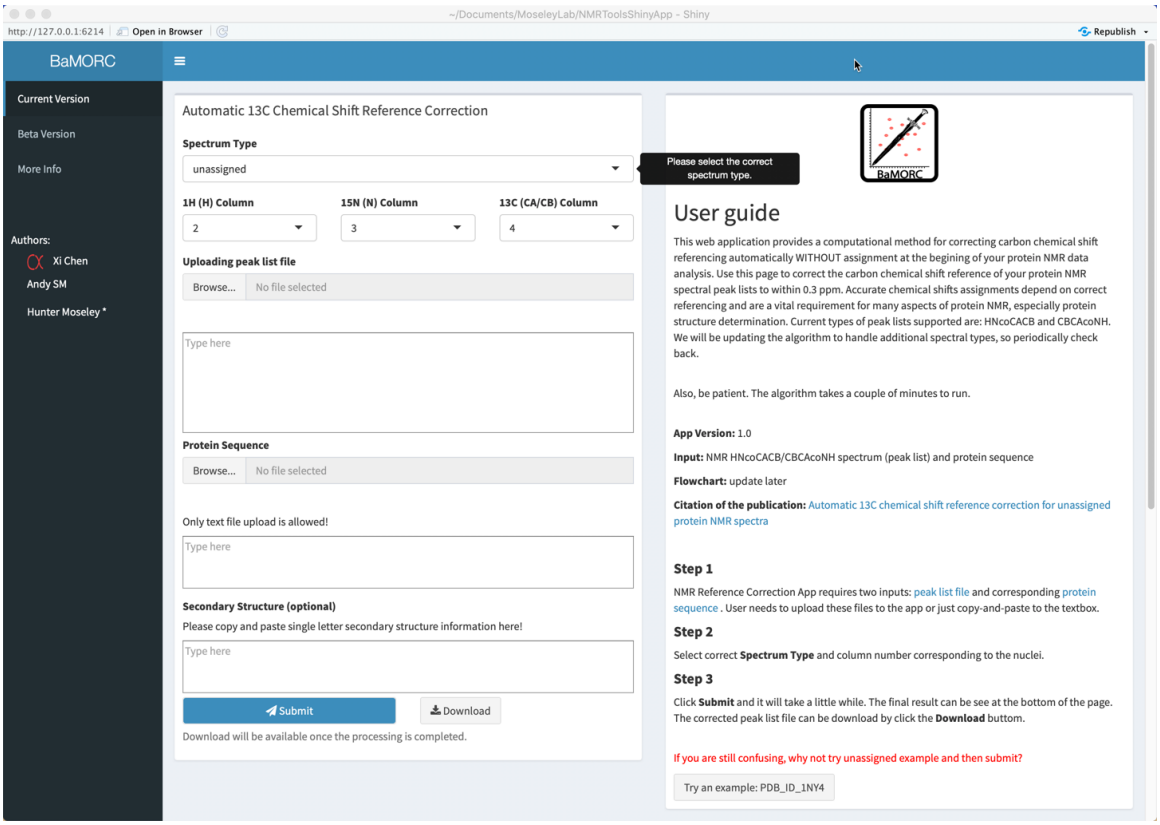

Figure 7.3 BaMORC web-based GUI landing page. Easy-to-use GUI allows researchers to use BaMORC methods and functions to reference correct assigned and unassigned NMR spectra.

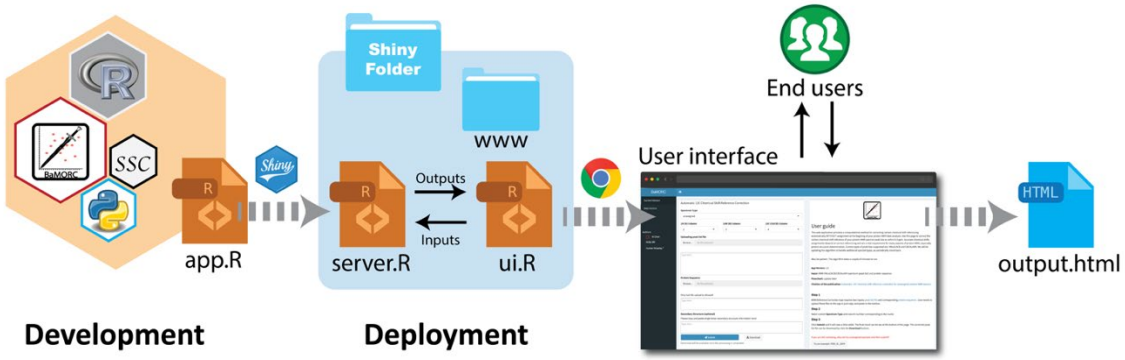

Figure 7.4 BaMORC web-based application implementation flowchart. After the development phase, the app.R utilizes the BaMORC package in the deployment phase to launch the web-based application or user interface. After the user supplies the input data, the web-based app will run the analysis and generate the report in html format.

7.4.4 BaMORC Shiny server app allows production-level integration.

To better serve the protein NMR community, we developed a BaMORC Shiny server app that provides the same web-based GUI but runs on a Shiny server, which can be integrated into web-based workflows (Figure 7.3). Similarly, as shown in Figure 7.4, the application folder that contains all of the necessary files could be directly put into the shiny app folder and can be automatically launched from the Shiny application server. This option is particular advantageous for NMR facilities and projects that provide web-based software for a user-base or to multiple sites. Use of a common online web-based workflow can improve reproducibility of results generated from multiple sites. Thus, this web app provides the advantage of avoiding potential discrepancies in data analysis that arise when proprietary or local methods are used in different laboratories. Relying on the Shiny package[136] and a container technology, either Docker or Singularity engines (Docker images are compatible with the Singularity engine) [132,133], the BaMORC Shiny server app is written in the open-source R programming language, distributed in a container as show in Figure 7.4 and Figure 7.5, and can be easily deployed. As in the standalone web-based GUI, datasets can be either copy-and-pasted or imported as a text or csv formatted file.

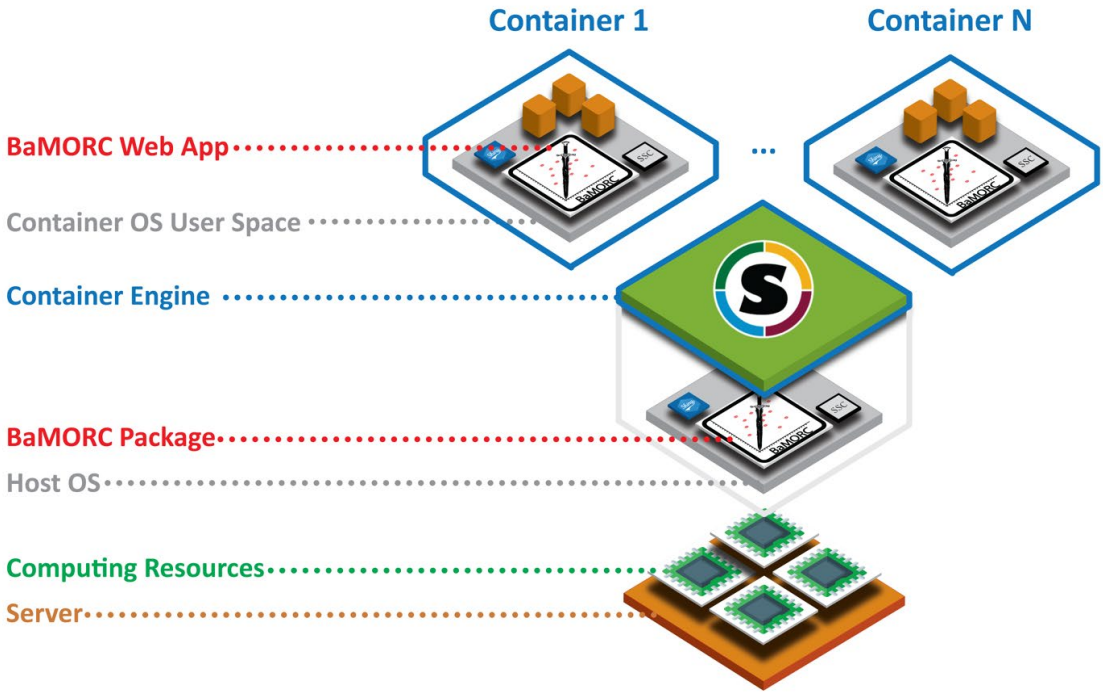

Figure 7.5 Production level integration through container technology. Through encapsulation of the OS system, library, and applications, BaMORC can be deployed in any research environment that supports the use of container technologies such as Docker and Singularity.

7.5 Discussion

BaMORC is a ^{13}C reference correction tool specifically designed for unassigned experimental peak lists and assigned sets of C_α and C_β resonances derived from protein samples. In a comparison with prior state of the art tool LACS, Assigned BaMORC achieves much better reference correction accuracy than the LACS algorithm. Based on a test set of over 500 RefDB datasets, Assigned BaMORC achieved a 0.40 ppm range for a 90% confidence interval, while LACS achieved a 0.59 ppm range for a 90% confidence interval. Also, the BaMORC method has robust performance even with 30% to 50% missing data. But beyond method performance, the implementation of BaMORC as an R package available in CRAN and on GitHub followed best software engineering principles to ensure software readability, maintainability, and reusability with strong methods

validation. The R package provides API, CLI, a web-based GUI, and a Shiny server app for integration into existing NMR data analysis pipelines and workflows. Using both unassigned and assigned BaMORC in an NMR data analysis workflow, a reference correction within +/- 0.2 ppm is achievable an estimated 90% of the time.

7.6 Conclusion

We have developed the BaMORC R package, which includes a new web-based GUI and Shiny server app, providing a protein NMR data preprocess tool that faces the increasing demand of reference correction without prior knowledge such as protein structure or assigned chemical shifts. Centered on simplicity, the web-based GUI is a standalone program that allows users with simply browser competence to perform reference correction using an intuitive web-based interface. The BaMORC Shiny App provides the same web-based interface but can be deployed on a Shiny server for easy incorporation into web-based NMR data analysis workflows. The software package and web app are freely available at:

https://cran.r-project.org/web/package/BaMORC/

CHAPTER 8. SUMMARY AND FUTURE DIRECTIONS

To summarize the research, one base algorithm with two applications, one software package, and one web application were designed and implemented in order to facilitate the protein carbon NMR reference correction either with or without assignment using Bayesian statistical model with statistics extracted from RefDB. The Bayesian Model Optimized Reference Correction algorithm was developed with many components. First, by adding covariance between alpha and beta carbon chemical shift, it allows a better statistical representation of the true chemical shift distribution and improves the statistical modeling. By representing cysteines as two types of amino acid based on its oxidation state, we further decreased mis-typing and improved the frequency estimation algorithm. By including the secondary structure information, we increased the typing from 19 amino acids to 57 compositions (frequencies) and improved the power of the optimization part of the project. To outperform prior state of the art algorithms, we also used an overlapping matrix as a representation of the statistical power of each classifier and transform the observer and predicted amino acid and secondary structure composition into an overlap-weighted composition. To improve the speed of the computation, we used a global optimization algorithm provided by the DEoptim, which reduced the run-time by at least two to three-fold as compared to the original grid-search approach.

One modification of the algorithm that could further improve the accuracy of the BaMORC is to use HNCACB NMR dipeptide chemical shifts information. In essence, BaMORC input data will have four features, they are two pairs of C_α and C_β chemical shifts from the sequential and intra-residue entries. However, in this approach, we might need to reconsider the covariance matrix implementation. As mentioned in the Chapter 7

that covariance between C_α and C_β chemical shifts from sequentially different experiments might not be a good estimation for the true covariance, similarly, the covariance between the sequential and intra-residue entries might not be well-captured either. One way to mediate the issue is to implement a 4×4 matrix with zero padding as following in Figure 8.1, however, the model might just statistically equivalent as the one using two 2×2 covariance matrix as the original BaMORC single peptide approach.

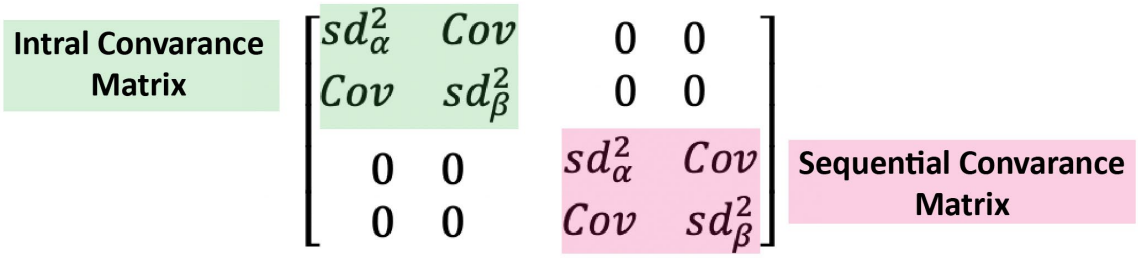

Figure 8.1 Dipeptide Covariance Matrix Implementation.

To further allow a broader use by the NMR community, we introduced a BaMORC R package (library), which includes an API and CLI, and a BaMORC web application. The library release allows the incorporation of the BaMORC functionality into an NMR data server, while the web application simplified the usage of the algorithm.

One of the most pertinent problems raised is determining deuteration levels of protein NMR samples to aid later research on protein structure and dynamics studies of complex biomolecules. A solution would recycle the existing algorithms, instead of adjusting the chemical shift values directly, the BaMORC algorithm would increase or decrease the deuteration level from 100% deuteration (perdeuteration) 0% and find the best deuteration level. The assumption for this solution is that the chemical shift affect from the deuteration is assumed to be uniform. Therefore, 1% deuteration level difference

will increase or decrease the affected chemical shift data by 1% of the associated sum of relevant deuteration shift effects.

We have shown that BaMORC can detect and correct ^{13}C chemical shift referencing errors before the protein resonance assignment step of analysis and without three-dimensional structure. By combining the BaMORC methodology with a new intra-peaklist grouping algorithm, we created a combined method called Unassigned BaMORC that utilizes only unassigned experimental peak lists and the amino acid sequence. Unassigned BaMORC kept all experimental three-dimensional HN(CO)CACB-type peak lists tested within +/- 0.4 ppm of the correct ^{13}C reference value. On a much larger unassigned chemical shift test set, the base method kept ^{13}C chemical shift referencing errors to within +/- 0.45 ppm at a 90% confidence interval. With chemical shift assignments, Assigned BaMORC can detect and correct ^{13}C chemical shift referencing errors to within +/- 0.22 at a 90% confidence interval. Therefore, Unassigned BaMORC can correct ^{13}C chemical shift referencing errors when it will have the most impact, right before protein resonance assignment and other downstream analyses are started. After assignment, chemical shift reference correction can be further refined with Assigned BaMORC. These new methods will allow non-NMR experts to detect and correct ^{13}C referencing error at critical early data analysis steps, lowering the bar of NMR expertise required for effective protein NMR analysis.

REFERENCES

1. Rabi, I. I. Space Quantization in a Gyrating Magnetic Field. *Phys. Rev.* **51,** 652–654 (1937).
2. Purcell, E. M., Torrey, H. C. & Pound, R. V. Resonance Absorption by Nuclear Magnetic Moments in a Solid. *Phys. Rev.* **69,** 37–38 (1946).
3. Bloch, F. Nuclear Induction. *Phys. Rev.* **70,** 460–474 (1946).
4. Saitô, H. Conformation-dependent 13C chemical shifts: A new means of conformational characterization as obtained by high-resolution solid-state 13C NMR. *Magnetic Resonance in Chemistry* **24,** 835–852 (1986).
5. Neal, S., Nip, A. M., Zhang, H. & Wishart, D. S. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR* **26,** 215–240 (2003).
6. Spera, S. & Bax, A. Empirical correlation between protein backbone conformation and C. alpha. and C. beta. 13C nuclear magnetic resonance chemical shifts. *J. Am. Chem. Soc.* (1991).
7. Wishart, D. S., Sykes, B. D. & Richards, F. M. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *Journal of Molecular Biology* **222,** 311–333 (1991).
8. Iwadate, M., Asakura, T. & Williamson, M. P. Cα and Cβ carbon-13 chemical shifts in proteins from an empirical database. *J. Biomol. NMR* (1999).
9. Wishart, D. S. & Case, D. A. *Use of chemical shifts in macromolecular structure determination.* (Methods in enzymology, 2001).
10. Mao, B., Guan, R. & Montelione, G. T. Improved Technologies Now Routinely Provide Protein NMR Structures Useful for Molecular Replacement. *Structure* **19,** 757–766 (2011).
11. Rosato, A. *et al.* Blind Testing of Routine, Fully Automated Determination of Protein Structures from NMR Data. *Structure* **20,** 227–236 (2012).
12. Serrano, P. *et al.* The J-UNIO protocol for automated protein structure determination by NMR in solution. *J. Biomol. NMR* **53,** 341–354 (2012).
13. Hoeck, C. R. in *Solving a 3D Structural Puzzle* **69,** 1–2 (Springer, Cham, 2018).
14. Gan, Z. *et al.* NMR spectroscopy up to 35.2 T using a series-connected hybrid magnet. *Journal of Magnetic Resonance* **284,** 125–136 (2017).
15. De Dios, A. C., Pearson, J. G. & Oldfield, E. Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. *SCIENCE-NEW YORK THEN …* (1993).
16. Vila, J. A. *et al.* Quantum chemical 13C(alpha) chemical shift calculations for protein NMR structure determination, refinement, and validation. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 14389–14394 (2008).
17. Akira, K., Hichiya, H., Shuden, M., Morita, M. & Mitome, H. Sample preparation method to minimize chemical shift variability for NMR-based urinary metabonomics of genetically hypertensive rats. *Journal of Pharmaceutical and Biomedical Analysis* **66,** 339–344 (2012).

18. Wu, J., An, Y., Yao, J., Wang, Y. & Tang, H. An optimised sample preparation method for NMR -based faecal metabonomic analysis. *Analyst* **135,** 1023–1030 (2010).
19. Vernon, R., Shen, Y., Baker, D. & Lange, O. F. Improved chemical shift based fragment selection for CS-Rosetta using Rosetta3 fragment picker. *J. Biomol. NMR* **57,** 117–127 (2013).
20. Lange, O. F. *et al.* Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 10873–10878 (2012).
21. Yang, S. & Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* (2010).
22. Barette, J., Velyvis, A., Religa, T. L., Korzhnev, D. M. & Kay, L. E. Cross-Validation of the Structure of a Transiently Formed and Low Populated FF Domain Folding Intermediate Determined by Relaxation Dispersion NMR and CS-Rosetta. *J. Phys. Chem. B* **116,** 6637–6644 (2011).
23. Khaneja, N., Reiss, T., Kehlet, C., Schulte-Herbrüggen, T. & Glaser, S. J. Optimal control of coupled spin dynamics: design of NMR pulse sequences by gradient ascent algorithms. *Journal of Magnetic Resonance* **172,** 296–305 (2005).
24. Meissner, A. & Sørensen, O. W. Sequential HNCACB and CBCANH Protein NMR Pulse Sequences. *Journal of Magnetic Resonance* **151,** 328–331 (2001).
25. James S Nowick, Omid Khakshoor, Mehrnoosh Hashemzadeh, A. & Brower, J. O. DSA: A New Internal Standard for NMR Studies in Aqueous Solution. *Org. Lett.* **5,** 3511–3513 (2003).
26. Ulrich, E. L. *et al.* BioMagResBank. *Nucl. Acids Res.* **36,** D402–8 (2008).
27. Wishart, D. S., Sykes, B. D. & Richards, F. M. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *Journal of Molecular Biology* **222,** 311–333 (1991).
28. Smelter, A., Rouchka, E. C. & Moseley, H. N. B. Detecting and accounting for multiple sources of positional variance in peak list registration analysis and spin system grouping. *J. Biomol. NMR* **68,** 281–296 (2017).
29. Grzesiek, S. & Bax, A. Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. *J. Am. Chem. Soc.* (1992).
30. Brillouin, L. A Theorem of Larmor and Its Importance for Electrons in Magnetic Fields. *Phys. Rev.* **67,** 260–266 (1945).
31. Arnold, J. T., Physics, M. P. T. J. O. C.1951. Variations in absolute chemical shift of nuclear induction signals of hydroxyl groups of methyl and ethyl alcohol. *aip.scitation.org*
32. Ramsey, N. F. & Purcell, E. M. Interactions between Nuclear Spins in Molecules. *Phys. Rev.* **85,** 143–144 (1952).
33. Physics, P. L. T. J. O. C.1957. C13 Nuclear Magnetic Resonance Spectra. *aip.scitation.org*

34. Anet, F., Society, A. B. J. O. T. A. C.1965. Nuclear Magnetic Resonance Spectral Assignments from Nuclear Overhauser Effects1. *ACS Publications*
35. Ernst, R. R. & Anderson, W. A. Application of Fourier Transform Spectroscopy to Magnetic Resonance. *Review of Scientific Instruments* **37,** 93–102 (2004).
36. LAUTERBER, P. C. Image formation by induced local interactions : examples employing nuclear magnetic resonance. *Nature* **246,** 469 (1974).
37. Aue, W. P., Bartholdi, E. & Ernst, R. R. Two-dimensional spectroscopy. Application to nuclear magnetic resonance. *The Journal of Chemical Physics* **64,** 2229–2246 (1976).
38. Jeener, J., Meier, B. H., Bachmann, P. & Ernst, R. R. Investigation of exchange processes by two-dimensional NMR spectroscopy. *The Journal of Chemical Physics* **71,** 4546–4553 (1979).
39. Wüthrich, K., Wider, G., Wagner, G. & Braun, W. Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *Journal of Molecular Biology* **155,** 311–319 (1982).
40. York, K. W. N.1986. *NMR of proteins and nucleic acids (book) Wiley-Interscience*.
41. Marion, D. *et al.* Overcoming the overlap problem in the assignment of proton NMR spectra of larger proteins by use of three-dimensional heteronuclear proton-nitrogen-15, 28 (15) pp 6150-6156 *Biochemistry*, (1989)
42. MESSERLE, B. A., Wider, G., Structural, G. O. N. I.1995. Solvent Suppression Using a Spin Lock in 2D and 3D NMR Spectroscopy with H2O Solutions. *World Scientific*
43. Kay, L. E., Ikura, M., Tschudin, R. & Bax, A. Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *Journal of Magnetic Resonance (1969)* **89,** 496–514 (1990).
44. Pervushin, K., Riek, R., Wider, G. & Wüthrich, K. Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci* **94,** 12366–12371 (1997).
45. Frericks, H. L., Zhou, D. H., Yap, L. L., Gennis, R. B. & Rienstra, C. M. Magic-angle spinning solid-state NMR of a 144 kDa membrane protein complex: E. coli cytochrome bo_3 oxidase. *J. Biomol. NMR* **36,** 55–71 (2006).
46. Kay, L. E., Ikura, M., Tschudin, R. & Bax, A. Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *Journal of Magnetic Resonance (1969)* **89,** 496–514 (1990).
47. Wishart, D. S. & Sykes, B. D. The ^{13}C Chemical-Shift Index: A simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J. Biomol. NMR* **4,** 171–180 (1994).
48. Cheung, M.-S., Maguire, M. L., Stevens, T. J. & Broadhurst, R. W. DANGLE: A Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. *Journal of Magnetic Resonance* **202,** 223–233 (2010).

49. Cornilescu, G., Delaglio, F. & Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13,** 289–302 (1999).
50. Shen, Y. *et al.* Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 4685–4690 (2008).
51. Wang, L., Eghbalnia, H. R., Bahrami, A. & Markley, J. L. Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. *J. Biomol. NMR* **32,** 13–22 (2005).
52. Wishart, D. S. & Case, D. A. Use of chemical shifts in macromolecular structure determination. *Methods Enzymol* **338,** 3–34 (2001).
53. Case, D. A., Dyson, H. J. & Wright, P. E. [13] Use of chemical shifts and coupling constants in nuclear magnetic resonance structural studies on peptides and proteins. *Methods Enzymol* **239,** 392–416 (1994).
54. Gronenborn, A. & Marius Clore, G. Identification of N-terminal helix capping boxes by means of 13C chemical shifts. *J. Biomol. NMR* **4,** (1994).
55. Metzler, W. J., Constantine, K. L., Friedrichs, M. S., Biochemistry, A. B.1993. Characterization of the three-dimensional solution structure of human profilin: proton, carbon-13, and nitrogen-15 NMR assignments and global folding pattern. *ACS Publications*
56. Wishart, D. S. & Sykes, B. D. [12] Chemical shifts as a tool for structure determination. *Methods Enzymol* **239,** 363–392 (1994).
57. Wishart, D. S. & Sykes, B. D. The ^{13}C Chemical-Shift Index: A simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J. Biomol. NMR* **4,** 171–180 (1994).
58. Harris, R. K., Becker, E. D., de Menezes, S. M. C., Goodfellow, R. & Granger, P. NMR nomenclature: nuclear spin properties and conventions for chemical shifts. IUPAC Recommendations 2001. International Union of Pure and Applied Chemistry. Physical Chemistry Division. Commission on Molecular Structure and Spectroscopy. *Magnetic Resonance in Chemistry* **40,** 489–505 (2002).
59. Nowick, J. S., Khakshoor, O., Hashemzadeh, M. & Brower, J. O. DSA: A New Internal Standard for NMR Studies in Aqueous Solution. *Org. Lett.* **5,** 3511–3513 (2003).
60. Wang, B., Wang, Y. & Wishart, D. S. A probabilistic approach for validating protein NMR chemical shift assignments. *J. Biomol. NMR* **47,** 85–99 (2010).
61. Chen, X., Smelter, A. & Moseley, H. N. B. Automatic 13C chemical shift reference correction for unassigned protein NMR spectra. *J. Biomol. NMR* **72,** 11–28 (2018).
62. Wishart, D. *et al.* 1H, 13C and 15N chemical shift referencing in biomolecular NMR. *J. Biomol. NMR* **6,** (1995).
63. Markley, J. L. *et al.* Recommendations for the presentation of NMR structures of proteins and nucleic acids. *Journal of Molecular Biology* **280,** 933–952 (1998).
64. Wishart, D. S. & Nip, A. M. Protein chemical shift analysis: a practical guide. *Biochemistry and Cell Biology* **76,** 153–163 (2011).

65. Wishart, D. S. *Interpreting protein chemical shift data*. (Progress in nuclear magnetic resonance spectroscopy, 2011).
66. Moseley, H. N. B., Sahota, G. & Montelione, G. T. Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. *J. Biomol. NMR* **28,** 341–355 (2004).
67. Wang, B., Wang, Y. & Wishart, D. S. A probabilistic approach for validating protein NMR chemical shift assignments. *J. Biomol. NMR* **47,** 85–99 (2010).
68. Ginzinger, S. W., Gerick, F., Coles, M. & Heun, V. CheckShift: automatic correction of inconsistent chemical shift referencing. *J. Biomol. NMR* **39,** 223–227 (2007).
69. Han, B., Liu, Y., Ginzinger, S. W. & Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR* **50,** 43–57 (2011).
70. Rieping, W. & Vranken, W. F. Validation of archived chemical shifts through atomic coordinates. *Proteins: Structure, Function, and Bioinformatics* **35,** n/a–n/a (2010).
71. Zhang, H., Neal, S. & Wishart, D. S. RefDB: a database of uniformly referenced protein chemical shifts. *J. Biomol. NMR* **25,** 173–195 (2003).
72. Grimsley, G. R., Scholtz, J. M. & Pace, C. N. A summary of the measured pK values of the ionizable groups in folded proteins. *Protein Science* **18,** 247–251 (2009).
73. Cojocari, D. *Amino Acids*.
74. Sibanda, B. L. & Thornton, J. M. β-Hairpin families in globular proteins. *Nature* **316,** 170–174 (1985).
75. Shafee, T. *Protein Secondary Structures*.
76. Prabakaran, S., Lippens, G., Steen, H. & Gunawardena, J. Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **4,** 565–583 (2012).
77. Bustamante, C., Cheng, W. & Mejia, Y. X. Revisiting the Central Dogma One Molecule at a Time. *Cell* **144,** 480–497 (2011).
78. Robinson, A. & van Oijen, A. M. Bacterial replication, transcription and translation: mechanistic insights from single-molecule biochemical studies. *Nature Reviews Microbiology 2013 11:5* **11,** 303–315 (2013).
79. Borrebaeck, C. A. K. Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. *Nature Reviews Cancer 2017 17:3* **17,** 199–204 (2017).
80. Landrieu, I. *et al.* NMR spectroscopy of the neuronal tau protein: normal function and implication in Alzheimer's disease. *Biochemical Society Transactions* **38,** 1006–1011 (2010).
81. Mukrasch, M. D. Untersuchung des Tau-Proteins mit Hilfe von NMR-Spektroskopie. (2007).
82. Nathalie Sibille *et al. Structural Impact of Heparin Binding to Full-Length Tau As Studied by NMR Spectroscopy†*. *Biochemistry* **45,** 12560–12572 (American Chemical Society, 2006).

83. Alain Sillen *et al.* NMR Investigation of the Interaction between the Neuronal Protein Tau and the Microtubules. *Biochemistry* **46,** 3055–3064 (2007).
84. Wang, Y. & Wishart, D. S. A simple method to adjust inconsistently referenced 13C and 15N chemical shift assignments of proteins. *J. Biomol. NMR* **31,** 143–148 (2005).
85. Han, B., Liu, Y., Ginzinger, S. W. & Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *Journal Biomolecular NMR* **50,** 43–57 (2011).
86. Zhang, H., Neal, S. & Wishart, D. Re-referenced Protein Chemical shift Database.
87. Allerhand, A., Gutowsky, H. S., Chemical, J. J. T. A.1966. Nuclear Magnetic Resonance Methods for Determining Chemical-Exchange Rates1. *ACS Publications*
88. Chen, G. & Balakrishnan, N. A General Purpose Approximate Goodness-of-Fit Test. *Journal of Quality Technology* **27,** 154–161 (2018).
89. Anderson, T. W. & Darling, D. A. A Test of Goodness of Fit. *Journal of the American Statistical Association* **49,** 765–769 (2012).
90. Lilliefors, H. W. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association* **62,** 399–402 (2012).
91. Salkind, N. *Chi-Square Test for Goodness of Fit*. (Sage Publications, Inc.). doi:10.4135/9781412952644.n78
92. Shapiro, S. S. & Francia, R. S. An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association* **67,** 215 (1972).
93. Sharma, D. & Rajarathnam, K. 13C NMR chemical shifts can predict disulfide bond formation. *J. Biomol. NMR* **18,** 165–171 (2000).
94. Fritzsching, K. J., Hong, M. & Schmidt-Rohr, K. Conformationally selective multidimensional chemical shift ranges in proteins from a PACSY database purged using intrinsic quality criteria. *J. Biomol. NMR* **64,** 115–130 (2016).
95. Wilson, E. B. & Hilferty, M. M. The distribution of chi-square. in (1931).
96. Krishnamoorthy, K., Mathew, T. & Mukherjee, S. Normal-Based Methods for a Gamma Distribution. *Technometrics* **50,** 69–78 (2012).
97. Drozdetskiy, A., Cole, C. & Procter, J. JPred4: a protein secondary structure prediction server. *Nucleic acids Research* (2015).
98. Cuff, J. A. & Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics* **40,** 502–511 (2000).
99. Goodfellow, I. NIPS 2016 Tutorial: Generative Adversarial Networks. *arxiv.org* (2016).
100. Mullen, K. M., Ardia, D., Gil, D. L., Windover, D. & Cline, J. DEoptim: An R Package for Global Optimization by Differential Evolution. **612,** (2009).
101. Mullen, K. M., Ardia, D., Gil, D. L., Windover, D. & Cline, J. DEoptim: An R package for global optimization by differential evolution. (2009).

102. Price, K., Storn, R. M. & Lampinen, J. A. *Differential evolution: a practical approach to global optimization*. (2006).
103. Ester, M., Kriegel, H. P., Sander, J., Kdd, X. X.1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *aaai.org*
104. Wenqing Feng, Roberto Tejero, Diane E Zimmerman, Masayori Inouye, A.Gaetano T Montelione. *Solution NMR Structure and Backbone Dynamics of the Major Cold-Shock Protein (CspA) from Escherichia coli: Evidence for Conformational Dynamics in the Single-Stranded RNA-Binding Site*. *Biochemistry* **37,** 10881–10896 (American Chemical Society, 1998).
105. Acton, T. B., Wu, M. J., Szyperski, T. & Montelione, G. T. Resonance assignments for the hypothetical protein yggU from Escherichia coli. *J. Biomol. NMR* (2003).
106. Moy, F. J., Seddon, A. P., Campbell, E. B. & Böhlen, P. 1 H, 15 N, 13 C and 13 CO assignments and secondary structure determination of basic fibroblast growth factor using 3D heteronuclear NMR spectroscopy. *Journal of biomolecular NMR* (1995).
107. Aramini, J. M. *et al.* Solution NMR structure of the 30S ribosomal protein S28E from Pyrococcus horikoshii. *Protein Science* **12,** 2823–2830 (2003).
108. Chien, C. Y., Tejero, R. & Huang, Y. A novel RNA-binding motif in influenza A virus non-structural protein 1. *Nature structural & molecular biology* (1997).
109. Sakurako Shimotakahara *et al.* NMR Structural Analysis of an Analog of an Intermediate Formed in the Rate-Determining Step of One Pathway in the Oxidative Folding of Bovine Pancreatic Ribonuclease A: Automated Analysis of 1H, 13C, and 15N Resonance Assignments for Wild-Type and [C65S, C72S] Mutant Forms†. *Biochemistry* **36,** 6915–6929 (1997).
110. Zheng, D., Aramini, J. M. & Montelione, G. T. Validation of helical tilt angles in the solution NMR structure of the Z domain of Staphylococcal protein A by combined analysis of residual dipolar coupling and NOE data. *Protein Science* **13,** 549–554 (2004).
111. Kelly A Mercier *et al. FAST-NMR: Functional Annotation Screening Technology Using NMR Spectroscopy*. *Journal of the American Chemical Society* **128,** 15292–15299 (American Chemical Society, 2006).
112. Schubert, M., Labudde, D., Oschkinat, H. & Schmieder, P. A software tool for the prediction of Xaa-Pro peptide bond conformations in proteins based on ^{13}C chemical shift statistics. *J. Biomol. NMR* **24,** 149–154 (2002).
113. *Statistics for Machine Learning*. (Packt Publishing Ltd.).
114. Opara, K. R. & Arabas, J. Differential Evolution: A survey of theoretical analyses. *Swarm and Evolutionary Computation* **44,** 546–558 (2019).
115. in *Contemporary Computer-Assisted Approaches to Molecular Structure Elucidation* 50–111 (Royal Society of Chemistry, 2011). doi:10.1039/9781849734578-00050
116. Venzke, J., Mascharka, D., Johnson, P., Davis, R. & Roth, K. Utilizing Machine Learning to Accelerate Automated Assignment of Backbone NMR Data. *analytics.drake.edu*

117. Emmons, J., Johnson, S., Urness, T. & Kilpatrick, A. Automated Assignment of Backbone NMR Data using Artificial Intelligence. *arXiv preprint arXiv:1506.05846* (2015).
118. Ginzinger, S. W., Gerick, F., Coles, M. & Heun, V. CheckShift: automatic correction of inconsistent chemical shift referencing. *J. Biomol. NMR* **39,** 223–227 (2007).
119. Chen, X., Smelter, A. & Moseley, H. N. Automatic 13 C chemical shift reference correction for unassigned protein NMR spectra. *J. Biomol. NMR* **72,** 11–28 (2018).
120. Smelter, A., Astra, M. & Moseley, H. N. B. A fast and efficient python library for interfacing with the Biological Magnetic Resonance Data Bank. *BMC Bioinformatics* **18,** 326 (2017).
121. Emmonsστ, J., Johnsonτ, S., Urness, T. & Kilpatrick, A. Automated Assignment of Backbone NMR Data using Artificial Intelligence. *johnemmons.com*
122. Moseley, H. N. B., Sperling, L. J. & Rienstra, C. M. Automated protein resonance assignments of magic angle spinning solid-state NMR spectra of β1 immunoglobulin binding domain of protein G (GB1). *J. Biomol. NMR* **48,** 123–128 (2010).
123. Michael C Baran, Yuanpeng J Huang, Hunter N B Moseley, A.Gaetano T Montelione. Automated Analysis of Protein NMR Assignments and Structures. *Chem. Rev.* **104,** 3541–3556 (2004).
124. Moseley, H., Monleon, D. & Montelione, G. T. Automatic determination of protein backbone resonance assignments from triple-resonance NMR data. *Methods in …* (2001).
125. Moseley, H. N. & Montelione, G. T. Automated analysis of NMR assignments and structures for proteins. *Current Opinion in Structural Biology* **9,** 635–642 (1999).
126. Chen, X., Smelter, A. & Moseley, H. N. B. Automatic 13C chemical shift reference correction for unassigned protein NMR spectra. *J. Biomol. NMR* **66,** 339 (2018).
127. *RStudio: Integrated Development for R. RStudio, Inc., Boston, MA; 2015*.
128. ca533, R. T. D.2018. R: A Language and Environment for Statistical Computing. (164AD).
129. van Rossum, G. & Drake, F. L. *The Python Language Reference Manual*. (Network Theory Ltd., 2011).
130. Chen, X., Smelter, A. & Moseley, H. N. B. *Bayesian Model Optimized Reference Correction (BaMORC) Method for Assigned and Unassigned Protein NMR Spectra*. (GitHub, 2018).
131. McIntosh, S., Kamei, Y., Adams, B. & Hassan, A. E. The impact of code review coverage and code review participation on software quality: a case study of the qt, VTK, and ITK projects. in 192–201 (ACM Press, 2014). doi:10.1145/2597073.2597076
132. Bernstein, D. Containers and Cloud: From LXC to Docker to Kubernetes. *IEEE Cloud Computing* **1,** 81–84 (2014).

133. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLoS ONE* **12,** e0177459 (2017).
134. Smelter, A., Astra, M. & Moseley, H. N. B. A fast and efficient python library for interfacing with the Biological Magnetic Resonance Data Bank. *BMC Bioinformatics* **18,** 326 (2017).
135. Chen, X. BMRBr: 'BMRB' File Downloader.
136. Shiny. *a web application framework for r (2015)*.

VITA

Xi (Bill) Chen

Education:

March 2019
Master (Certificate) in Statistics - University of Kentucky, Lexington, KY

May 2005
Bachelor in English – Tianjin University of Technology, Tianjin, China

Working Experiences

April 2018 – 2019
Deep Learning Institute (DLI) Certified Instructor
Deep Learning Institute, NVIDIA

August 2013 – March 2019
Graduate Research Assistant
Prof. Derek S. Young, Dept. of Statistics, University of Kentucky, Lexington, KY

August 2017 – Feb 2019
Research Collaborator
Moseley Bioinformatics Lab, University of Kentucky, Lexington, KY

August 2011 – May 2012
Teaching Assistant
Dept. of Chemistry, University of Louisville, Louisville, KY

August 2010 – May 2011
Lab Assistant
Prof. Aaron T. Setterdahl, Dept. of Chemistry, Indiana University Southeast, New Albany, IN

June 2005 – May 2008
International Coordinator / Event Organizer
Tianjin Culture Bureau, Tianjin, China

Publication

- **Parallelized Interactive Machine Learning on Autonomous Vehicles,** NAECON Dec 2018 DOI: 10.1109/NAE- CON.2018.8556776. Link
- **A Method to Facilitate Cancer Detection and Type Classification from Gene Expression Data using a Deep Autoencoder and Neural Network**, arXiv, Dec 2018 arXiv:1812.08674. Link
- **Deep Learning by Doing: The Nvidia Deep Learning Institute**, Journal of Computational Science Education, Dec 2018 DOI: 10.22369/issn.2153-4136/10/1/16. Link

- **Pan-Cancer Epigenetic Biomarker Selection from Blood Sample Using SAS(R)**, MWSUG, Sep 2018. Link
- **Automatic 13C Chemical Shift Reference Correction for Unassigned Protein NMR Spectra**, Journal of Biomolecular NMR, Aug 2018 DOI: 10.1007/s10858-018-0202-5. Link
- **Finite Mixture-of-Gamma Distributions: Estimation, Inference, and Model-Based Clustering**, Expected 2019.

CERTIFICATION

- SAS Certified Clinical Trials Programmer: Link
- Deeplearn.ai Certification Series (AI/DL/RL/ML): Link
- SAS Certified Statistical Business Analyst: Link
- SAS Certified Advanced Programmer: Link
- DataCamp R Programming Certifications
- NVidia Certifications: Link
- GCP/AWS Certifications: Link
- NVidia Certifications: Link

AWARDS

- AAAI Student Poster Traval Award - 2019
- MWSUG Conference Poster Scholarship - 2016/2018
- SIAM SDM Dissertation Traval Award - 2018
- Grow with Google Challenge Scholarship - 2018
- 3rd Venture Competition UK - 2017
- Azure Research Funding - 2016
- MWSUG Conference Poster Scholarship - 2016/2018